EMPIRICAL STUDIES IN HOSPITAL EMERGENCY DEPARTMENTS

Robert Johnson Batt

A DISSERTATION

in

Operations and Information Management

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

*Signature*_____

Christian Terwiesch, Andrew M. Heller Professor of Operations and Information Management

Graduate Group Chairperson

*Signature*_____

Eric Bradlow, Professor of Marketing, Statistics, and Education

Diwas S. KC, Assistant Professor of Information Systems & Operations Management, Emory University

Sergei Savin, Associate Professor of Operations and Information Management

Olanrewaju A. Soremekun, Assistant Professor of Emergency Medicine, Hospital of the University of Pennsylvania

EMPRICIAL STUDIES IN HOSPITAL EMERGENCY DEPARTMENTS

Robert Johnson Batt

to Michelle, Annika, and Peter

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Christian Terwiesch for teaching me so much over the last five years. He has helped me develop my research taste, skill, and style. I have learned many things from him about how to navigate and succeed in the academic world. Also, he was my advocate and staunch supporter through the long job-market process. He has also provided funding for me to attend several conferences and for much of my fifth-year support. I am grateful for the lessons learned and the support received.

Thanks also go to my dissertation committee, Professor Diwas Kc, Professor Sergei Savin, and Doctor Olan Soremekun. They provided much guidance and insight over the years. Doctor Jesse Pines gave me my first glimpse inside an emergency department and has been a collaborator and mentor since the first year of the program. He is always full of energy and good ideas.

Thanks to the OPIM faculty and staff. This has been an excellent department to be a part of and I will always be proud to be a part of the OPIM family. I look forward to attending reunions at INFORMS each year for many years to come. I am indebted to the Wharton Risk Management and Decisions Processes Center and the Fishman-Davidson Center for Service and Operations Management for providing additional funding for my research and for my fifth year.

I want to especially thank the three men who have served as OPIM PhD coordinators

during my time here: Professors Serguei Netessine, Gerard Cachon, and Noah Gans. Serguei recruited an excellent cohort and helped us all see what a great place Wharton is. Gerard successfully got us through the first year and the qualifier. Noah has been carrying us along since then and has done an excellent job of helping us pick out our path through the program. He has always been willing to run defense for us when needed. I am grateful for his straight-forward advice.

Speaking of the cohort, I cannot imagine a more wonderful group with whom to have shared this journey. We have all had moments of excitement and moments of frustration, but it has been a great comfort and joy to share these all with Alison, Jaelynn, Joel, and Santiago. I will dearly miss the lunches and swapping of stories. I want to give special thanks to Santiago for humbly playing "delivery boy" for so many lunches and for being my companion on the ride home many many days. I will miss those conversations.

We lost one member of our cohort along the way. Brent Dooley was surely the most likely of us all to succeed in the Operations Management world, and he helped me greatly with many of the first-year courses. It is a shame he had to drop out of the program, but it is much to his credit that he sacrificed his own career to care for an ailing family member.

Beyond my own cohort, the other doctoral students of the OPIM department have been great friends and colleagues as well. I learned much from those who went before me and I have likewise enjoyed sharing with those behind me. I hope that OPIM continues to be a place where the doctoral students help each other, enjoy each other, and laugh together.

I am grateful for those friends I've made here in Philadelphia apart from work. Our dear friends from Church of the Saviour in Wayne, PA mean more to us than I can say. I have been edified by them and have enjoyed many many laughs and meals with them. Of all things in Philadelphia, I will miss them most.

I give my greatest thanks to my family. I am thankful for those who've gone before me that helped make this possible. Our extended family–parents, siblings, aunts and uncles, and

cousins–have been a great encouragement and also an endless source of help and babysitting at home. Annika has filled my office with artwork and my days with smiles. Peter, who was born during the program, is my best buddy. I am thankful that they have put up with a dad who has had such a strange job.

Finally, my amazing wife, Michelle: She has sacrificed more than anyone to make this possible. I am forever grateful for her selfless acts of love in caring for the kids, keeping the house running, and putting up with the craziness of these five years. More importantly, I am grateful for the way she has challenged me when I needed it and encouraged and believed in me over and over on the tough days. Thank you!

> "For I know the plans I have for you," declares the LORD, "plans to prosper you and not to harm you, plans to give you hope and a future."
>
> Jeremiah 29:11

ABSTRACT

EMPRICIAL STUDIES IN HOSPITAL EMERGENCY DEPARTMENTS

Robert Johnson Batt

Christian Terwiesch

This dissertation focuses on the operational impacts of crowding in hospital emergency departments. The body of this work is comprised of three essays. In the first essay, *"Waiting Patiently: An Empirical Study of Queue Abandonment in an Emergency Department,"* we study queue abandonment, or "left without being seen." We show that abandonment is not only influenced by wait time, but also by the queue length and the observable queue flows during the waiting exposure. We show that patients are sensitive to being "jumped" in the line and that patients respond differently to people more sick and less sick moving through the system. This study shows that managers have an opportunity to impact abandonment behavior by altering what information is available to waiting customers. In the second essay, *"Doctors Under Load: An Empirical Study of State-Dependent Service Times in Emergency Care,"* we show that when crowded, multiple mechanisms in the emergency department act to retard patient treatment, but care providers adjust their clinical behavior to accelerate the service. We identify two mechanisms that providers use to accelerate the system: early task initiation and task reduction. In contrast to other recent works, we find the net effect of these countervailing forces to be an increase in service time when the system is crowded. Further, we use simulation to show that ignoring state-dependent service times leads to

modeling errors that could cause hospitals to overinvest in human and physical resources. In the final essay, *"The Financial Consequences of Lost Demand and Reducing Boarding in Hospital Emergency Departments,"* we use discrete event simulation to estimate the number of patients lost to Left Without Being Seen and ambulance diversion as a result of patients waiting in the emergency department for an inpatient bed (known as boarding). These lost patients represent both a failure of the emergency department to meet the needs of those seeking care and lost revenue for the hospital. We show that dynamic bed management policies that proactively cancel some non-emergency patients when the hospital is near capacity can lead to reduced boarding, increased number of patients served, and increased hospital revenue.

Contents

List of Tables

List of Figures

CHAPTER 1 : Introduction

This dissertation focuses on the operational impacts of crowding in hospital emergency departments (EDs). A typical ED encounter progresses through five stages: arrival, triage, waiting, service, and discharge or boarding. Following a survey of the ED operations literature, the body of this work is comprised of three essays which address each of these stages.

The first essay, *"Waiting Patiently: An Empirical Study of Queue Abandonment in an Emergency Department,"* focuses on the waiting phase and how crowding affects the queue abandonment behavior of patients waiting for treatment. Despite a large literature on customer satisfaction in queues that suggests that people are more tolerant of waiting when they are kept informed of why and how long they must wait (e.g., Hui and Tse 1996), in most EDs in America, patients are given little or no information about how long they will be required to wait for service. Patients are expected to wait patiently until called for service. Hospitals may be reluctant to share queue status information for fear of "sticker shock" causing patients to abandon quickly. However, a utility theory view of customer behavior suggests that customers make a stay or abandon decision by weighing the benefit of obtaining service against the expected cost of continuing to wait. A major element of that decision is the customer's estimate of the remaining wait time. The practice of providing no information does not take into account the fact that patients can partially observe the queue, and that they likely make wait time estimates based on what they see. We use patient-level time stamp data to reconstruct both the queue size and the associated arrival and departure flows observed by each patient during their waiting experience. Using this data, we find that patients that observe either high queue census or arrivals to the queue exhibit increased likelihood of abandoning. In contrast, observing a high rate of departures into the service area leads to a lower abandonment probability. We also find that patients infer the relative health status of those around them and respond differently to the movement of patients that are relatively more sick or less sick. All of these effects are consistent with patients adjusting their

wait time estimate based on what they observe. Because patient abandonment behavior is affected by observable queue events, managers have an opportunity to affect abandonment behavior by manipulating the information available to patients.

The second essay, *"Doctors Under Load: An Empirical Study of State-Dependent Service Times in Emergency Care,"* focuses on the triage and service phases of the ED encounter. Specifically, we examine how the service process changes in response to crowding in the ED. Classic queuing theory assumes service times to be independent of the state of the system, and recent Operations Management literature has shown evidence of service times of worker-paced systems decreasing with workload. However, in the ED, we observe that service times exhibit an inverted-U relationship with system load, first increasing then decreasing as the queue grows. We develop a theoretical framework of Speedup and Slowdown effects that gives rise to this nonmonotonic relationship. We show that isolated elements of the ED service process, such taking an x-ray or receiving a medication, take longer during periods of high load. This works to lengthen the total service time of the ED encounter. However, we find that care providers (doctors and nurses) alter their clinical behavior to speed up the service time. When the ED is crowded, triage nurses tend to order more diagnostic tests that otherwise would be ordered by the doctor later on in the encounter. This early task initiation reduces the service time by allowing some activities, such as lab processing, to happen while the patient is in the waiting room rather than in a treatment bed. The other clinical change is that for some subsets of patients, doctors reduce the overall amount of diagnostic testing ordered, which also reduces service time. The combination of these Slowdown and Speedup responses gives rise to the inverted-U relationship of service time and system load. Thus we show that complex service systems can indeed be state-dependent, in contrast to classic queuing theory, and that the state-dependency need not be a simple monotonic relationship as observed in recent Operations Management studies.

The third essay, *"The Financial Consequences of Lost Demand and Reducing Boarding in Hospital Emergency Departments,"* examines the connection between "boarding" and lost

demand. Boarding is the practice of holding patients in the ED while they wait to be transferred to an inpatient bed in the hospital. In busy hospitals, boarding patients may wait for twelve or more hours for an inpatient bed (Carr et al. 2010). Several medical studies have shown that this has negative clinical consequences for patients (e.g., Carr et al. 2007a, Chalfin et al. 2007). From an operational perspective, boarding is problematic because it reduces the effective capacity of the ED. This can lead to increased levels of patients abandoning the queue due to long waits. Further, some hospitals cope with ED congestion by diverting incoming ambulances to other hospitals. Both abandonment and diversion represent a failure of the hospital to serve all potential customers and to receive the associated revenue. Thus, reducing boarding times would lead to lower lost demand and higher revenues. Using discrete event simulation, we estimate the impact of reducing mean boarding time by one hour and find that it leads to serving an additional 4.2 patients per day, or approximately $12,000 of revenue per day. However, there is a downside to reducing boarding. Serving more patients in the ED creates an increase in inpatient bed demand as some of the ED patients are admitted to the hospital. If non-ED (elective) patients have to be bumped from the schedule to accommodate the increase in ED-admitted patients, the hospital loses money since, on average, non-ED admitted patients generate more revenue per day than do patients admitted from the ED. However, the hospital is rarely running at capacity and thus non-ED patients do not always have to be bumped. We show that active bed management policies that proactively bump scheduled non-ED patients only when the hospital approaches capacity allows the hospital to serve the increase in ED-admit demand without sacrificing revenue. In summary, this paper shows the extent to which boarding leads to lost ED demand and that recapturing this demand through efforts to reduce boarding times can be financially beneficial for hospitals.

CHAPTER 2 : Literature Review

While each essay of this dissertation provides a detailed review of the literature relevant for the specific topic, we provide here a broad overview of the literature of emergency department operations in general. Since our intent is for our work to contribute to both the Operations Management (OM) and the Emergency Medicine (EM) management communities, this survey draws from both of bodies of work.

We first propose a conceptual framework of ED operations to help organize the following survey. In the OM literature, EDs are most commonly treated as queuing systems in much the same way call centers are treated as queuing systems. Call centers have been extensively studied (c.f. Gans et al. 2003), and we borrow concepts for our framework from such works as Brown et al. (2005) which breaks call center operations into an arrival process, a waiting/abandonment process, and a service process.

The most influential framework of ED operations from the EM literature is the Input-Throughput-Output model proposed by Asplin et al. (2003). The Input-Throughput-Output model is a conceptual model of ED operations that was developed to help the EM community address the problem of ED crowding. Drawing liberally from queuing theory concepts, the Input-Throughput-Output essentially breaks ED operations into drivers of demand (input), drivers of treatment time (throughput), and barriers to discharging patients (output).

We propose a different and somewhat more detailed decomposition of ED operations than Asplin et al. (2003). We view ED operations as being comprised of four segments: Demand Generation, Front Side Operations, Back Side Operations, and Disposition & Departure.

2.1. Demand Generation

Demand for ED services can be categorized in several ways. Asplin et al. (2003) divide demand into three categories based on clinical need: emergency care, urgent unscheduled

care, and safety net care. This categorization is closely related to the triage level classification systems used to prioritize patients upon arrival. The three levels of Asplin et al. (2003) generally correspond to high, medium, and low priority conditions. We refer to this categorization as defining the "type" of care.

Another way to categorize demand is by whether the demand arrives exogenously or endogenously. We refer to this as the "source" of care. Most ED demand is generated exogenously from the community at large and from the overall health care system. Andersen and Laake (1987) provides a helpful behavioral model of healthcare demand generation that is driven by factors such as patient need for healthcare services, predisposing factors that affect an individual's likelihood of seeking care, and enabling factors that affect an individual's ability to access care.

A portion of ED demand is endogenously generated in that it is a result of the ED operations or decisions made during care. For example, patients that abandon the ED queue (left without being seen) frequently return within a day or two to seek care, and some return in a more severe condition requiring increased care than had they been treated upon first arrival (Baker et al. 1991, Rowe et al. 2006). Similarly, patients who are discharged too quickly may have to return for additional care (Derlet et al. 2001). This revisit or "bounce back" phenomenon has also been observed with intensive care units and transitional care units in hospitals (Chan et al. 2012, Kc and Terwiesch 2012). Revisits can also occur if a discharged patient has poor access to appropriate follow-up care through a primary care physician or other ambulatory care provider (Rask et al. 1994).

A third classification of demand is by arrival mode. The majority of ED arrivals are "walkin" arrivals, or patients that come by their own means of transportation. At our study hospital, over 70% of arrivals are walkin arrivals, while nationally about 75% of arrivals are walkins (Niska et al. 2010). Ambulance arrivals account for the next largest arrival mode (approximately 25% in the study hospital). The remainder is made up of arrival modes such as helicopter, police, and other public services.

Arrival mode is operationally relevant for two reasons. First, walkin and ambulance arrivals go through different pre-treatment processes. Walkin arrivals go through a multi-step check-in, triage, and registration process generally followed by a waiting period before beginning treatment. Ambulance arrivals, in contrast, generally skip this pre-treatment process and get moved to a treatment bed quickly, regardless of clinical need. The other reason arrival mode matters, is that for many hospitals, ambulance arrivals can be diverted to other hospitals, thereby giving the hospital the ability to partially control the arrival rate to the ED. In contrast, federal law mandates that all walkin arrivals be provided, at a minimum, a medical evaluation and stabilizing treatment.

Ambulance diversion is a controversial topic. Theoretically, diversion is a operational mechanism that helps pool the medical resources of a community to serve the public, but the reality of it tends to fall short (Deo and Gurvich 2011). Diversion has also been shown to lead to longer ambulance transport times which can lead to worse clinical outcomes (Schull et al. 2003b). Operational factors such as ED size, inpatient utilization, and number of boarding patients in the ED have all been found to affect the use of ambulance diversion (Deo et al. 2013, Schull et al. 2003a).

Each of these demand categorizations (type, source, arrival mode) provide a different view of the arriving patients, and each categorization is operationally useful since the patient routing and required services is potentially different for each type.

2.2. Front Side Operations

Front side operations include all processes and actions that occur before the patient is moved to a treatment bed (the back side of the ED). Front side operations tend to be focused on walkin arrivals since other arrivals usually come in a separate entrance and skip many of the front side steps. In most EDs, the two basic front side processes are triage and waiting. The triage process involves a nurse making a brief assessment of the patient and assigning a triage score. The patient then waits until called for service. There is also a registration

process (providing an address, insurance information, etc.) that occurs at some point during each ED encounter. This is usually done while the patient is waiting. There are many ways to accomplish these front side tasks, and there are many variations on these basic tasks that have been studied both in the OM and EM literature. See Wiler et al. (2010) for an excellent survey of the relevant EM literature.

The main purpose of the triage process is to assign a triage score. In the United States, the most common triage system is the five-level Emergency Severity Index (ESI) system (Baumann and Strout 2005, Gilboy et al. 2011), but other systems are also in use (Storm-Versloot et al. 2011). The triage score is an indication of clinical acuteness and is generally used as a priority classification to determine the order in which patients are served. Saghafian et al. (2013) have proposed a triage system that also takes into account the expected complexity of treating the patient. Thus, their system is essentially a new application of the well-known $c\mu$-rule for priority queues (Wolff 1989). Argon and Ziya (2009) use the ED triage setting as motivation to explore the problem of how to assign priorities under imperfect information.

The triage score does not strictly define the order in which patients are served because many hospitals are now employing separate "tracks" for different categories of patients. The most common being a FastTrack that serves low-acuteness, quick-service-time patients (Meislin et al. 1988). FastTracks generally make use of dedicated physical space and care providers in order to disconnect from the more complex workings of the regular ED.

Several studies have shown FastTracks to be quite effective at reducing the length of stay in the ED for the target population (O'Brien et al. 2006, Nash et al. 2007). FastTracks have been so effective that some hospitals are now testing adding an additional MidTrack to serve patients of mid-level acuteness but with conditions that have well defined and straightforward care (Soremekun et al. 2012, Urgent Matter Learning Network II 2010). At least one hospital has attempted creating treatment tracks based on the probability of a patient needing to be admitted (King et al. 2006). Saghafian et al. (2012) examine this type of tracking and find that it can be beneficial in EDs with a high percentage

of admitted patients and with long boarding times. While the idea of creating separate service tracks may seem to be counter to the standard "pooling is better" lessons taught in introductory Operations Research texts (Hillier and Lieberman 2010), several recent papers have shown that partitioning customers and servers is optimal when there are multiple types of customers with different service requirements (Whitt 1999a, Ata and Van Mieghem 2009, Hu and Benjaafar 2009).

Another modification to the standard triage process that has been explored is early initiation of testing or treatment. In our study hospital triage nurses are given authority to order several different types of diagnostic tests such as a urinalysis, a basic blood test, or a simple x-ray. Another common adaptation is what is known as "standing orders" or "advances triage protocols." Standing orders allow a triage nurse to order a predefined set of tests or procedures if a patient meets a set of criteria (Campbell et al. 2004, Cooper et al. 2008). The operational rationale behind early initiation of care is that tasks that would otherwise be done with the patient in a treatment bed are accomplished while the patient is in the waiting room. This reduces the amount of time the patient spends in the treatment bed, which is frequently the bottleneck resource. We discuss this topic in greater detail in 4.

A related approach for achieving early initiation of care is to add a physician to the triage process. This provides two main benefits. First, similar to implementing standing orders, tests and treatments can be started much earlier in the patient encounter, but without the limitations placed on triage nurses. Second, for some patients, the physician can provide all necessary care at the triage station and immediately discharge the patient (Soremekun et al. 2011). This is not a possibility for the triage nurse because federal law requires that each patient be seen by a higher level care provider before being discharged. This immediate treatment and discharge dramatically reduces the patient length of stay and eliminates the need for a treatment bed for the given patient. Several studies had been published on the use of a physician at triage with most showing improvements in time until medical evaluation and total length of stay (e.g., Rogers et al. 2004, Choi et al. 2006).

The triage process as described above is sometimes modified based on ED busyness. During busy times, some hospitals position a greeter nurse as the first point of contact with an arriving patient before triage (Weber et al. 2011, Rogg et al. 2013). The greeter nurse does a very quick medical assessment to determine if immediate care is needed, and collects basic information, such as name and age, to start a medical chart. The greeter nurse can also set the order in which patients should be seen by the triage nurse. In essence, the greeter nurse performs a pre-triage triage. At the other extreme, when there is no queue and treatment beds are available, some hospitals forgo front side operations altogether and move arriving patients directly into treatment beds. This is known as direct bedding. This is usually accompanied by bedside registration, where the registration personnel come to the treatment bed to collect the necessary information (Bertoty et al. 2007, Takakuwa et al. 2007).

The other major activity that is part of the front side operations is waiting. After triage, patients wait to be called for service. However, some patients choose to abandon the queue before they are called for service. This is referred to as "left without being seen" (LWBS). While the required wait time is almost certainly a factor in the stay or abandon decision, other factors, such as the patient's condition and the crowd level in the waiting room may also factor into the decision. Queue abandonment is the focus of Chapter 3 and we direct the reader there for a detailed review of the related literature.

2.3. Back Side Operations

We define back side operations as all the processes and actions taken to diagnose, treat, and disposition a patient once the patient is placed in a treatment bed. This is also referred to as the service or treatment phase of the ED visit (Batt and Terwiesch 2013). There are many facets to back side operations and they can be analyzed from several perspectives. At one level, the treatment phase can be viewed as a black-box process that requires a random amount of processing time. For example, (Batt et al. 2013) examines how the throughput of the ED back side is affected by the number of in-service and boarding patients.

The process can also be analyzed from the patient point of view wherein the patient is a "job" moving through a queuing network, not unlike a jobshop (Jackson 1963). From this vantage point, one is interested in not only total time in system but also the path through the system and the amount of processing time and waiting time. This viewpoint is also helpful when considering the amount of pooling or partitioning of servers as described in Section 2.2.

Back side operations analysis can also focus on the service resources, such as doctors, nurses, and equipment, rather than the patient. Issues such as staffing levels (Green et al. 2006b, 2012) and work load allocation (Armony and Ward 2010) are two key issues. Another topic of interest is the "what to do next?" question faced by doctors and nurses. These care provides are responsible for several patients at a time and at the completion of each task they must decide which task to do next. Saghafian et al. (2012) shows that the optimal decision rule depends on whether the care provider is trying to minimize total time or some short-term objective such as time to first order. Similarly, Dobson et al. (2012) shows that if in-process patients generate work by way of interruptions, then care providers should prioritize serving patients near the end of treatment and keep the the number of in-process patients low.

Much recent work has focused on how the state of the system (i.e., busyness) affects the service rate of the various resources. For example, Kc and Terwiesch (2009) finds that hospital transport personnel work faster when the workload is high. Kc (2012) examines physician multi-tasking and shows that productivity has an inverted-U relationship with the level of multi-tasking. The state-dependent nature of service times in the ED is the focus of Chapter 4, and we direct the reader there for a detailed review of the related literature.

The EM literature on back side operations (excluding clinical/medical methods) is largely focused on reporting the results of process improvement projects. For example, studies have examined the benefits of new technologies such as point-of-care testing (Singer et al. 2005, Jang et al. 2013), electronic patient tracking (Boger 2003, Aronsky et al. 2008), and personal

communication devices (Le et al. 2004, Walsh and Yamarick 2005). Interestingly, there is little agreement in the EM community about exactly how to measure ED performance, and thus it is hard to compare the results of process improvements. Hwang et al. (2011) presents a comprehensive survey of more than 2,600 studies related to crowding and identifies over 70 different crowding or performance related metrics used in the studies. Thus, defining ED performance metricsis an important first step in future work.

2.4. Disposition & Departure

The term disposition is used as both a noun and a verb in the ED. The noun form refers to a patient's destination after leaving the ED, usually either admitted to the hospital or discharged to go home. The verb form refers to the act of deciding not only the post-ED destination, but also when the patient is ready to leave the ED for that destination. Dispositioning a patient is generally the final decision the physician must make regarding a patient and it signals the end of diagnosis and treatment in the ED. Patients that are discharged depart the ED soon after being dispositioned. Patients that are admitted, however, frequently have to wait in the ED for some time. These patients are referred to as boarders (Asplin et al. 2008).

While the disposition decision is mainly driven by the patient's medical condition, there are operational factors that also must be accounted for by the physician. The decision is complex because the state of the system at the time of disposition affects the disposition decision, and the disposition decision affects the future state of the system. For example, if there are no available inpatient beds, an admitted patient may have to board for several hours, and this has been shown to have a negative affect on clinical outcomes for some types of patients (e.g., Carr et al. 2007b, Singer et al. 2011). The boarding patient continues to occupy a treatment bed and thus reduces the effective capacity of the ED to serve other patients. However, a patient that is inappropriately discharged may return later for more care and may be in a worse condition (Shiber 2010).

The disposition decision is conceptually similar to a two-tier gatekeeper system where the gatekeeper must decide whether to serve the customer directly or to pass the customer along to a higher-skilled, more expensive resource (Shumsky and Pinker 2003). Hasija et al. (2005) shows that when each tier is modeled as an M/M/N queue there is a critical level of customer complexity above which all customers should be referred to the second tier. Likewise, motivated by security checkpoints at border crossings, Zhang et al. (2011) finds a similar optimal referral policy that balances waiting costs and misclassification costs.

The routing of patients from the ED to inpatient beds is also an interesting problem. In most hospitals, the inpatient beds are divided into wards that specialize in certain conditions (e.g., ICU, cardiac care, orthopedics, general medicine, etc.). Thus ED doctors not only consider if any bed is available, but if the right kind of bed is available. Thompson et al. (2009) formulate this problem as a finite-horizon Markov decision process and achieve optimal bed allocation by proactively transferring some inpatients between wards to make room for newly arriving patients. Mandelbaum et al. (2012) examine a similar bed allocation problem but further considers fairness in work allocation between the wards.

Boarding has received a great deal of attention in the EM community. It is widely viewed as one of the major causes of ED crowding (Rabin et al. 2012). Boarding time has been shown to be correlated with both the number of people in the ED and the utilization level of inpatient beds (McCarthy et al. 2009). Despite the fact that boarding consumes ED resources and has a deleterious effect on clinical outcomes, it is still a common occurrence. This has led some to suggest that hospitals tolerate boarding because it implicitly frees up beds for more profitable elective patients (Mitka 2008). Chapter 5 explores this topic and finds that boarding is not revenue enhancing once the increased abandonment and diversion caused by boarding are taken into account.

Having presented an overview of ED operations and the relevant areas of research, we now turn to the three essays that comprise the bulk of this dissertation.

CHAPTER 3 : Waiting Patiently: An Empirical Study of Queue Abandonment in an Emergency Department[1]

3.1. Introduction

The body of knowledge on queuing theory is voluminous and spans almost a century of research. However, one of the least understood aspects of queuing theory is human behavior in the queue. Understanding the human element is crucial in designing and managing service-system queues such as quick-serve restaurants, retail checkout counters, call centers, and emergency departments.

Specifically, queue abandonment (also known as reneging) is one aspect of human behavior that is poorly understood. Abandonment is undesirable in most service settings because it leads to a combination of lost revenue and ill-will. In a hospital emergency department, abandonment takes on the added dimension of the risk of a patient suffering an adverse medical event. While the hospital may not be legally responsible for such an event, it is certainly an undesirable outcome.

Prior literature has explored psychological responses to waiting and has generally found that people are happier and waiting seems less onerous when people are kept informed of why they are waiting and how long the wait will last (Hui and Tse 1996). Given these findings, it seems almost trivial that it is beneficial to provide waiting customers with as much information as possible about the wait. In practice, however, many service systems, such as call centers and emergency departments, which provide limited or no information to waiting customers. One reason for this is that uninformed customers might naively estimate the waiting time to be short and thus join a queue which they would not join if they were informed about the expected waiting time. Sharing information with customers about the queue status is an active area of analytical queuing theory research (e.g. Armony et al.

[1]This chapter is based on Batt, Robert J., Christian Terwiesch. 2013 "Waiting Patiently: An Empirical Study of Queue Abandonment in an Emergency Department." Working Paper.

2009, Plambeck and Wang 2012). Yet, there exists limited empirical work studying how queue status information affects customers. An exception to this is the recent work by Lu et al. (2012), which provides evidence that even in a simple queuing system in which all information is fully observable and customers are served in their order of arrivals, customers might not use the available information rationally.

The empirical setting of our work is a hospital emergency department (ED). In this setting, waiting patients can observe the waiting room but they cannot observe the service-delivery portion of the system (the treatment rooms). Additionally, even though patients can observe the waiting room, it is not at all clear what they can learn from what they observe. Factors such as arrival order, priority level, assignment to separate service channels, and the required service time of others are not readily apparent. Interestingly, most American EDs provide no queue-related information to the patients. The position of the American College of Emergency Physicians is that providing queue information might have "unintended consequences" and lead to patients who need care leaving without treatment (ACEP 2012). However, this position does not account for how patients respond to the information they do have: what they see.

In this paper, we focus on how what patients observe and experience over the course of the waiting exposure impacts their abandonment decisions. Using detailed timestamp data of 180,000 patient visits that we obtained from the ED's electronic patient tracking system, we are able to reconstruct a set of variables that patients should rationally have considered in their decision whether to abandon the queue when they were in the waiting room. Our theoretical framework hypothesizes that patients observe and consider two types of variables, stock variables and flow variables. Stock variables are those that describe the number of other patients in the waiting room, such as the total number of patients, the total number of patients with a higher priority, or the total number of patients with a later arrival time. Flow variables are those that describe the rate with which the queue is depleted as well as the rate with which new patients arrive, such as the number of arrivals in the last hour, the

number of departures in the last hour, or the number of patients that have been served in the last hour before patients who had an earlier arrival time. Some of these variables can be directly observed by the patient, while others have to be inferred. For example, the number of patients in the waiting room is directly observable to the patient, while, given that the priority data is not shared with all patients, the number of patients in the waiting room with a high priority score can only be inferred. This novel approach towards predicting and estimating abandonment behavior of ED patients allows us to make the following four contributions:

1. We find that for patients of moderate severity, observing an additional patient in the queue increases the probability of abandonment by half a percentage point, even when appropriately controlling for wait time. This is equivalent to a 15 minute increase in wait time and extends the prior result of Lu et al. (2012) from a deli counter to an emergency room.

2. We show that the observed flow of patients in and out of the waiting room has an effect on abandonment, with arrivals leading to increased abandonment and departures leading to decreased abandonment. Given the unknown priority of newly arriving patients, the patients in the waiting room are more likely to abandon the queue when new patients arrive after them, as they fear being overtaken by these new arrivals. Regarding departures, we show that patients respond differently to outflows that maintain first-come-first-served order and those that do not. For example, observing an additional waiting room departure that maintains first-come-first-served order reduces the probability of abandonment by 0.6 percentage points, equivalent to a 19 minute reduction in wait time. In contrast, observing an additional waiting room departure that violates first-come-first-served has an insignificant impact on abandonment.

3. We show that patients respond to more than just the "facts" that they observe. They make inferences about the severity of other patients and respond differently to the flow of more and less severe patients. For example, we find that observing an additional

arrival of a patient sicker than oneself increases the probability of abandonment by one percentage point whereas observing the arrival of a patient less sick than oneself has no discernible effect on abandonment. Further, we show that patients are quite adept at making these relative severity inferences.

4. We show that early initiation of a service task, such as diagnostic testing, reduces abandonment. For example, receiving an order for a diagnostic test during the triage process reduces the probability of abandonment by 1.8 percentage points. This is particularly interesting because unlike the other variables examined in this paper, early service initiation does not impact the waiting time.

These contributions show that patient abandonment behavior is affected by the waiting patients experience while in the waiting room. Thus, a queue is not either visible (like in a grocery store) or invisible (like in a call center), but often times combines aspects of both. In such settings, providing no information to customers does not mean that customers are without queue information. Further, to the extent the visual queue information is misleading or does not lead to the desired behavior, managers have an opportunity to intervene by altering what information is available to the patients. For example, providing separate waiting rooms for different triage levels would reduce abandonment due to observing a crowded waiting room and due to obscuring arrivals of higher priority patients.

3.2. Clinical Setting

Our study is based on data from a large, urban, teaching hospital with an average of 4,700 ED visits per month over the study period of January, 2009 through December, 2011. The ED has 25 treatment rooms and 15 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. 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 operates somewhat autonomously from the rest of the

ED in that it utilizes seven dedicated beds and is usually staffed by a dedicated group of Certified Registered Nurse Practitioners rather than Medical Doctors.

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.

The study hospital operates in a manner similar to many hospitals across the United States (Batt and Terwiesch 2013). Upon arrival, patients are checked in by a greeter 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 official chief complaint. The triage nurse assigns a triage level, which indicates acuteness, using the five-level Emergency Severity Index (ESI) 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 diagnostic tests, for example an x-ray or a blood test. Patients are generally not informed of their assigned triage level nor are they given any queue status information.

After triage, patients wait in a single waiting room to be called for service. Patients are in no way visibly identified, thus a waiting patient does not know what triage level other patients have been assigned. Further, patients can sit anywhere in the waiting room, thus there is no ready visual signal of arrival order. There is no queue status information posted in the waiting room.

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 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. For example, the ED might serve a triage level 4 patient if the patient has been waiting a long time and there are not more acute patients that need immediate attention. Similarly, the FT might serve a triage level 3 patient if the patient has been waiting a long time and the patient's needs can be met by the nurse practitioners in the FT.

Most patients likely have little or no understanding that the ED and FT coexist and work as separate service channels. Further, since patients go through the same doors to begin service in either the ED or the FT, there is no visual indication to the remaining waiting patients as to which service channel a patient has been assigned.

Once a patient is called for service, a nurse escorts the patient to a treatment room and the treatment phase of the visit begins. When treatment is complete, the patient is either admitted to the hospital or discharged to go home. If a patient is not present in the waiting room when called for service, that patient is temporarily skipped and is called again later, up to three times. If the patient is not present after a third call, the patient is considered to have abandoned, the patient record is classified as Left Without Being Seen (LWBS), and is closed out. The time until a record is closed out as LWBS is usually quite long, with a mean time of over four hours (about triple the mean wait time for those who remain). Note that a patient is free to abandon the ED at any time. However, for this study, we focus solely on abandonment that occurs before room placement.

3.3. Literature Review

The classical queuing theory approach to modeling queue abandonment is the Erlang-A model first introduced by Baccelli and Hebuterne (1981). In the Erlang-A model, each customer has a maximum time she is willing to wait, and she waits in the queue until she either enters service or reaches her maximum wait time, at which point she abandons the queue. The maximum wait times are usually assumed to be i.i.d. draws from some distribution,

commonly the exponential (Gans et al. 2003). Examples of work using the Erlang-A model include Brown et al. (2005) and Mandelbaum and Momcilovic (2012). Modeling abandonment in this way provides analytical tractability, but does not shed light on the actual drivers of customer behavior.

An alternative view of queue abandonment is based on customer utility maximization. In such models, customers are assumed to be forward-looking and balance the expected reward from service completion against the expected waiting costs. Thus, there are generally three terms of interest in these models: the reward for service, the instantaneous unit waiting cost, and the estimated residual waiting time (Mandelbaum and Shimkin 2000, Aksin et al. 2012). Some models also include a discount rate, which adds a fourth term of interest.

One of the key findings from this body of literature is that abandoning the queue is not rational in many M/M/c type queues (Hassin and Haviv 2003). However, since this conclusion does not match well with observation of real queuing systems, there is a rich literature of studies which modify the basic queue model to generate rational abandonments. For example, Haviv and Ritov (2001) and Shimkin and Mandelbaum (2004) consider the case of nonlinear waiting costs leading to abandonment. Mandelbaum and Shimkin (2000) considers customer abandonment from a system with a possible "fault state" in which service will never be initiated. Such a fault state can occur in an overloaded multi-class queue, such as in an ED. If the arrival rate of high priority customers is large enough, the queue becomes unstable for low priority customers and the wait goes to infinity (Chan et al. 2011). See Hassin and Haviv (2003) for a review of assumptions that lead to rational abandonments.

Another possibility is that customers are boundedly rational, meaning that there is some error in their estimation of the cost of waiting. Bounded rationality has been studied in several settings, as reviewed by Gino and Pisano (2008). Huang et al. (2012) examines how bounded rationality affects the queue joining decision and Kremer and Debo (2012) finds evidence of bounded rationality in queue joining in laboratory experiments. To the best of our knowledge, bounded rationality has not been studied in regard to queue abandonment.

A related avenue of active queuing research addresses queues with various levels of information. Much of this work is motivated by the call-center industry and determining what information a call center should provide to its customers. For example, Guo and Zipkin (2007) compare M/M/1 queue performance when no, partial, and full information is revealed. They find that providing information always either improves throughput or customer utility, but not necessarily both. Similarly, Jouini et al. (2009) and Armony et al. (2009) both examine the impact of delay announcements on abandonment behavior in multi-server, invisible queues and find that providing more information can improve system performance with little customer loss. Plambeck and Wang (2012) shows that if customers exhibit time-inconsistent preferences through hyperbolic discounting, then hiding the queue may be welfare maximizing while being suboptimal for the service provider.

The question of what to tell waiting customers has also been explored. Many papers have focused on developing wait time estimators under various queuing disciplines that can be used to provide customers credible information (e.g, Whitt 1999b, Ibrahim and Whitt 2011). Given an estimated wait time distribution, Jouini et al. (2011) explores what value from the wait time distribution should be provided to the customer to balance the customers' balking probability with the provider's desire for high throughput. Allon et al. (2011) considers the "what the to tell customers" question under the assumption of strategic behavior by both customers and providers.

There are many studies from a variety of fields that identify drivers of queue abandonment. While they generally do not explicitly mention the three terms of the utility function, they can be mapped to this framework to aid in understanding their contributions and differences. For example, Larson (1987) discusses such issues as perceived queue fairness and waiting before or after service initiation, both of which likely impact expected residual time. Janakiraman et al. (2011) studies the psychological phenomena of goal commitment and increasing "pain" of waiting which are equivalent to increasing service reward and increasing waiting costs respectively in the utility framework. Bitran et al. (2008) provides a survey of

other such findings from the marketing and behavioral studies domains.

The medical literature contains several empirical studies on drivers of abandonment from emergency departments. Demographic factors (e.g., age, income, and race), institutional factors (e.g., hospital ownership and the presence of medical residents), and operational factors (e.g., utilization level) have all been shown to influence patient abandonment (Hobbs et al. 2000, Polevoi et al. 2005, Pham et al. 2009, Hsia et al. 2011).

While there are several recent empirical Operations Management papers dealing with queuing systems in the healthcare setting (e.g., Batt and Terwiesch 2013, Berry Jaeker and Tucker 2012, Chan et al. 2012), none have focused on queue abandonment. There are, however, two recent papers that study queue abandonment empirically, one in a call center and one at a deli. Aksin et al. (2012) uses a structural model to estimate the underlying service reward and waiting cost values for customers calling into a bank call center. Under assumptions of an invisible queue, linear waiting costs, and known exogenous hazard functions, the study finds that customers are heterogeneous in their parameter values and that ignoring the endogenous nature of abandonment decisions may lead to misleading results in various queuing models. Our work differs from Aksin et al. (2012) in terms of both setting and methodology. Our study setting is a semi-visible, multi-class queue (in the ED, the waiting room is visible but the clinical treatment area is not) as compared to an invisible multi-class queue. In terms of methodology, to estimate the latent structural parameters, Aksin et al. (2012) imposes strong structural assumptions (e.g. known common hazard function, linear waiting costs, past time is sunk, etc.). In contrast, we are not estimating any structural parameters and thus we use reduced form models which require fewer structural assumptions.

Lu et al. (2012) examines how aspects of a visible queue, such as queue length and number of servers, affect customer purchase behavior at a grocery deli counter. One of the key findings of this paper is that customers are influenced by line length but are largely immune to changes in the number of servers, even though the number of servers has a large impact on wait time. Stated differently, customers are boundedly rational in that they do not

appropriately incorporate all available information into their balk or abandon decisions.

Our work differs from Lu et al. (2012) in several ways. First, our setting is more complex. Lu et al. (2012) examines a fully visible, single-class, FCFS queue as compared to the semi-visible, multi-class queue in the ED. Second, because our data are richer and more detailed than in Lu et al. (2012) we are able to examine a broader set of questions regarding queue behavior and we can do so with fewer inferences about the customer experience. For example, Lu et al. (2012) must infer if a customer observed the queue, when a customer observed the queue, and what was the length of the queue observed. In contrast, our data allow us to know both when a patient entered the queue and what was the queue length at that moment. Further, we observe the dynamics of the queue during the waiting experience including arrivals, departures, and the patient mix. Thus we are able to not only confirm the key result of Lu et al. (2012) regarding queue length, but we are also able to examine how the observed flow and fairness of the queue impacts the abandonment decision. Thus, we believe our work serves to expand the understanding of the behavior of customers waiting in line.

3.4. Framework & Hypotheses

The primary purpose of this study is to determine to what extent the visible aspects of the queue impact the abandonment decision. In the ED, just because the hospital does not provide queue status information does not mean that the patients are completely without queue status information. Patients can observe the number of people in the waiting room and the flow of patients in and out of the waiting room. Understanding the impact of these visual cues on abandonment will help identify possible ways to influence abandonment behavior by manipulating the information available to waiting patients. We intentionally do not address the issue of whether abandonment is good or bad. That depends on the hospital's objective function and defining that is beyond the scope of this paper. However, we provide a few thoughts on the issue in the discussion section of the paper (Section 3.9).

We now develop a theory of how patients respond to visible queue elements. Abstracting from the optimal stopping problem formulation of Aksin et al. (2012), we assume that the abandonment decision is the result of a patient repeatedly evaluating the following personal utility function:

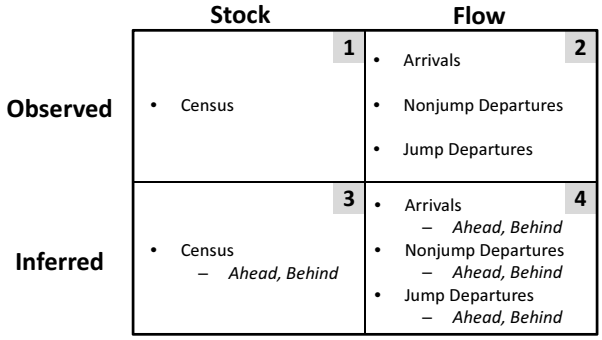
$$\text{Utility} = \max\left[\mathbb{E}\left[\begin{pmatrix}\text{Service}\\ \text{Reward}\end{pmatrix} - \begin{pmatrix}\text{Wait}\\ \text{Cost}\end{pmatrix} \times \begin{pmatrix}\text{Residual}\\ \text{Wait}\end{pmatrix}\right], 0\right] \tag{3.1}$$

The service reward is the utility gained from receiving treatment. The wait cost is the disutility incurred for each unit of wait time. The residual wait is the time remaining until service is commenced. While all three terms of the utility function may have some uncertainty or may change over the course of the waiting exposure, we are most interested in the formation of the expected residual wait time as this is the term that is most clearly affected by the queue evolution. Any information that increases the expected residual wait will increase the probability of the patient abandoning. Also following Aksin et al. (2012), we assume that past waiting costs are sunk and are irrelevant for future decisions.

Given that the hospital provides no information regarding the residual wait, the waiting experience itself is the only source of information that should impact the residual wait estimate. We categorize the visible queue information into four classes of variables created by the permutations of two pairs of classifications: stocks and flows, and observed and inferred (Figure 1). The key "stock" of interest is the waiting room census, while the key "flows" are the arrivals and departures from the waiting room. By "observed" and "inferred" we mean that some things can be objectively observed, such as the number of arrivals to the ED, while others can only be inferred, such as the number of patients in the waiting room with a higher triage classification than one's own.

Quadrant 1 of Figure 1 contains the only observed stock variable: Census. This waiting room census is the first, and perhaps most salient, visual cue that a waiting patient ob-

Figure 1: Visible Queue State Variables



serves. If patients behave according to the Erlang-A model, such that wait time is the only determinant of abandonment, then waiting room census should have no impact on abandonment, controlling for wait time. However, if patients behave in a utility maximizing way, then increasing waiting room census likely increases the patient's residual time estimate and abandonment probability (Guo and Zipkin 2007). This leads to our first hypothesis:

Hypothesis 1 Controlling for wait time, abandonment increases with waiting room census.

This relationship between census (queue length) and queue balking/abandoning behavior is the focus of Lu et al. (2012). We compare our results with Lu et al. (2012) in the sequel.

Quadrant 2 lists the observed flow variables: Arrivals and two types of Departures (nonjump and jump, defined below). At our study hospital, arrivals and departures are quite easy to observe if a patient chooses to do so. There is a single entry door for walk-in patients, and there is a single door that leads into the clinical treatment area. If the ED were a pure first-come first-served (FCFS) system, then one would expect arrivals to have little or no effect on abandonment. However, since the ED is a priority-based system, new arrivals may well jump the line and be served before currently waiting patients. Therefore, arrivals may cause waiting patients to adjust their residual time estimate upward leading to more abandonment.

Hypothesis 2A Abandonment increases with observed arrivals.

We define departures from the waiting room to include only departures to begin treatment (we address abandonments later). Patients that observe a high departure rate may take this as a signal that the system is moving quickly and therefore adjust their residual time estimate downward, leading to less abandonment. However, if a departure is a "jump," that is Patient A arrives before Patient B but Patient B enters service before Patient A, then this provides a mixed signal to Patient A. It signals system speed, which presumably reduces the residual time estimate. However, the jump departure does not move Patient A any closer to service, and thus the reduction in residual time estimate is less than for a regular departure. There may also be a psychological effect on Patient A if Patient A views the jump as unfair. This would increase the (psychological) waiting cost in the utility function and cause Patient A to be more likely to abandon. These possibilities lead to the following two hypotheses.

Hypothesis 2B Abandonment decreases with observed departures.

Hypothesis 2C Jump departures decrease abandonment less than nonjump departures.

Note that what we refer to as a "jump" is equivalent to what Larson (1987) terms a "slip" and Whitt (1984) terms "overtaking."

The above hypotheses consider the patient response to observable stock and flow variables. We now consider how patient inferences might modify behavior. While patients may not have a full understanding of the ED queuing system, they are likely aware that the ED operates on a priority basis rather than a FCFS basis. In fact, there are multiple placards in the waiting room explaining this point. Thus, patients may recognize that the presence of sicker patients can impact their wait time differently than less sick patients. However, since all patient information is kept confidential, patients can only infer the relative priority of those around them in the waiting room. Certainly, this is an inexact process at best, but likely not a pointless endeavor.

As we consider the variables shown in Quadrants 3 and 4, we want to determine if patients are able to differentiate between those who are ahead of and behind them in the priority queue

and if this affects their behavior. While we leave the precise definitions of the Quadrant 3 and 4 variables to Section 3.5, the general principle is that each variable is split into two parts. One part measures those who are ahead in line according to the priority queue scheme and the other part measures those who are behind the given patient according to the priority queue scheme. A fully informed, rational patient would respond only to those ahead of them in the queue since those behind them should not impact the patient's wait time. For example, observing a larger number of patients in the waiting room of equal or higher priority than an arriving patient (Census Ahead) should increase abandonment (assuming Hypothesis 1 is true) while the number of people of lower priority (Census Behind) should have no effect on abandonment at all. However, since patients can only infer the priority of others, they may make some classification errors and react to those behind them in the queue. Therefore we state our hypotheses in terms of comparing the effects of the ahead and behind variables.

Hypothesis 3 Abandonment increases more with the census of those ahead in the priority queue than it does with the census of those behind in the priority queue.

Hypothesis 4A Abandonment increases more with arrivals of those ahead in the priority queue than it does with arrivals of those behind in the priority queue.

Hypothesis 4B For departures that maintain arrival order (nonjump departures), abandonment decreases more with departures of those ahead in the priority queue than it does with those behind in the priority queue

Hypothesis 4C For departures that violate arrival order (jump departures), abandonment decreases more with departures of those ahead in the priority queue than it does with those behind in the priority queue.

For each of these four preceding hypotheses, the null hypothesis is that the effect of the ahead and behind variables is equal. This would occur if patients are unable to reliably distinguish the relative queue position of the other waiting patients.

While the above hypotheses focus on visual queue elements impacting the expected residual wait time and hence the abandonment behavior, another factor that potentially impacts the residual wait time estimate is the patient experience. Specifically, early initiation of diagnostic testing at triage may influence abandonment. Being assigned a test by the triage nurse may lead to a patient perceiving herself as being of relatively high priority and thus having a lower residual wait time. There could also be a psychological effect, as hypothesized by Maister (1985), that the perception of wait time is shorter once the patient perceives service to have started. This leads to our final hypothesis:

Hypothesis 5 Abandonment decreases with triage testing.

3.5. Data Description, Definitions, & Study Design

We now describe the dataset and define the key variables. In the discussion below, the index t indicates an 15-minute interval in the study period, the index T indicates the patient triage level, and the index i denotes a patient visit to the ED, not a specific patient. Note that some patients do have multiple visits, and we control for this with clustered standard errors (described in detail in Section 3.6). Further, because we estimate all models for each triage class separately, the index i is actually an index within the triage class.

Our data include patient level information on over 180,000 patient visits to the ED including demographics, clinical information, and timestamps. Patient demographics include age, gender, and insurance classification (private, Medicare, Medicaid, or none). Clinical information includes pain level on a 1 to 10 scale (10 being most severe), chief complaint as recorded by the triage nurse, and a binary variable indicating if the patient had any diagnostic tests, such as labs or x-rays, ordered at triage. Timestamps include time of arrival, time of placement in a treatment room, and time of departure from the ED. Table 1 provides descriptive statistics of the patient population by triage level. We do not study ESI 1 patients because these patients do not abandon. However, we do include ESI 1 patients in all relevant census measures in the analysis.

Table 1: Summary Statistics

	ESI 2	ESI 3	ESI 4	ESI 5
Age	49.8	39.0	34.7	34.2
	(0.11)	(0.07)	(0.07)	(0.14)
%Female	54%	66%	58%	51%
	(0.003)	(0.002)	(0.002)	(0.005)
Pain (1-10)	4.5	5.5	5.4	4.1
	(0.03)	(0.02)	(0.02)	(0.04)
%FastTrack	2%	3%	68%	67%
	(0.001)	(0.001)	(0.002)	(0.005)
Wait Time(hr.)	1.0	1.9	1.3	1.3
	(0.01)	(0.01)	(0.01)	(0.01)
Service Time(hr.)	3.7	4.0	1.8	1.2
	(0.02)	(0.01)	(0.01)	(0.01)
Census at Arrival	13.9	11.7	11.9	11.4
	(0.06)	(0.04)	(0.05)	(0.09)
%LWBS	1.7%	9.5%	4.7%	7.4%
	(0.001)	(0.001)	(0.001)	(0.003)
N	27,538	65,773	39,878	10,509

Means shown. Standard error of mean in parentheses

Empirical analysis on customer abandonment is often confounded by censored or missing data. Ideally, one would observe each customer's willingness to wait and the actual wait time if she stayed. However, only the minimum of these two is ever realized (actual wait time or actual abandonment time), leading to censored data. In the study hospital, abandonment times are not observed, leading to missing data for all patients who abandon. We know neither when they left, nor how long their wait would have been had they stayed for service. We address this missing data problem in two ways. In Section 3.7.1 we follow Zohar et al. (2002) and take averages across time to estimate the system waiting time. In Section 3.7.2 we use the wait times of similar patients who arrived in temporal proximity to create an estimated offered wait time for those who abandon.

For the regression models, we are interested in how the *offered wait time* impacts the abandonment decision. The offered wait is the wait time had the patient remained for service (Mandelbaum and Zeltyn 2013). For patients who do remain, this is their actual wait ($WAIT_i$), which we calculate directly from the timestamps. For patients who abandon, we

must estimate their offered wait ($\widehat{WAIT}_i$). We do this by calculating the average of the wait times of the two chronologically adjacent patients (one before and one after) who did not abandon . To get a sense of the accuracy of the estimated offered wait time $\widehat{WAIT}_i$, we examine the deviation between $\widehat{WAIT}_i$ and $WAIT_i$ for all patients that did not abandon. The deviation has a mean of 0.00 and a standard deviation of 1.1 hours. 50% of the values are are between $\pm$0.3hours, and more than 80% of the values are between $\pm$1hour. Thus, $\widehat{WAIT}_i$ appears to be unbiased, and is relatively close to the true value.

We then define the offered wait time as follows

$$OWAIT_i = \begin{cases} WAIT_i & \text{if patient stays} \\ \widehat{WAIT}_i & \text{if patient abandons} \end{cases} \tag{3.2}$$

To calculate the waiting room census measure, we divide the study period into 15-minute intervals labeled t, and we use the patient visit timestamps to generate the census variable $INTERVAL\_CENSUS_t$ as the number of patients in the waiting room during interval t. We also decompose the census measure into the waiting room census of each of the five ESI triage classes ($INTERVAL\_CENSUS_{t,T}$, $T \in \{1,2,3,4,5\}$). We assign a census value to each patient ($CENSUS_i$) based on the time of arrival. For example, for patient i who arrives at time interval t, $CENSUS_i = INTERVAL\_CENSUS_t$. We likewise create the variable $BEDS_i$ as the number ED treatment beds in use, which is the number of patients in the treatment phase of the visit.

In order to test Hypothesis 3, we would ideally decompose $CENSUS_i$ into those patients whom patient i perceives to be more sick and less sick than herself. However, since these perceptions are not observed by the econometrician, we proxy for them by using the triage classification of the waiting patients to calculate the census of those ahead of and behind patient i assuming a priority queue system without preemption that serves patients on a FCFS basis within a priority level. Therefore, any waiting patient of equal or higher priority (lower

ESI number) is considered as ahead of the arriving patient ($CENSUS\_AHEAD_i$), and any waiting patient of lower priority (higher ESI number) is considered as behind the arriving patient ($CENSUS\_BEHIND_i$). We emphasize that these variables are defined for each patient relative to the given patient's own triage level. For example, for an ESI 3 patient, patients in the waiting room of ESI levels 1 through 3 are counted in the $CENSUS\_AHEAD_i$ variable and patients of ESI levels 4 and 5 would be counted in the $CENSUS\_BEHIND_i$ variable.

The flow variables needed to test Hypotheses 2A,B,C and 4A,B,C are constructed based on the patient timestamps. For each patient visit we calculate the number of arrivals ($ARRIVE_i$) and departures ($DEPART_i$) that occur within one hour of patient i's arrival. Further, we create alternative departure variables $NONJUMP_i$ and $JUMP_i$ based on whether the departing patient(s) arrived before or after patient i respectively. As with the census variable, we also decompose the flow variables by triage level ($ARRIVE_{i,T}$, $DEPART_{i,T}$, $NONJUMP_{i,T}$, $JUMP_{i,T}$, $T \in \{1, 2, 3, 4, 5\}$).

We split each flow variable into two parts as follows based on those ahead and behind the given patient according to the priority queuing scheme.

- $ARRIVE\_AHEAD_i$: Arriving patients with higher priority than patient i
- $ARRIVE\_BEHIND_i$: Arriving patients with equal or lower priority than patient i
- $DEPART\_AHEAD_i$: Departing patients with equal or higher priority than patient i
- $DEPART\_BEHIND_i$: Departing patients with lower priority than patient i
- $NONJUMP\_AHEAD_i$: Departing patients with equal or higher priority than patient i and that arrived before patient i
- $NONJUMP\_BEHIND_i$: Departing patients with lower priority than patient i and

that arrived before patient i

- $JUMP\_AHEAD_i$: Departing patients with higher priority than patient i and that arrived after patient i

- $JUMP\_BEHIND_i$: Departing patients with equal or lower priority than patient i and that arrived after patient i

Note that the jump/nonjump language indicates relative arrival timing only, while the ahead/behind language indicates relative position in the priority queue which is a function of both arrival timing and priority level.

Once we add these flow variables to the model, we must restrict the sample to those who have been in the system some moderate amount of time to allow for observation of the system flow. Specifically, we restrict the sample to only patients with an offered wait of greater than one hour. Since the flow variables just described ($ARRIVE_i$, $DEPART_i$, $NONJUMP_i$, $JUMP_i$, etc.) are defined as the flows during the first hour after arrival of patient i, we are effectively asking the question, "what is the effect of flow during the first hour on patients who stay at least an hour," rather than the more broad ideal question of, "how does observed flow affect abandonment?" This sample restriction reduces the sample size by about half, and makes a significant finding less likely.

When we restrict the sample to patients with an offered time of greater than one hour it is possible that those who abandon do so quickly and are not actually in the waiting room for an hour to observe the flows. However, if this is the case, this should bias our results toward the null hypothesis of flow variables having no effect since patients who abandon quickly would not observe many arrivals or departure. Thus, any significant results are likely conservative estimates of the impact of the flow variables.

3.6. Econometric Specification

We now develop the econometric specifications for testing our hypotheses. Since we are studying the behavior of individuals making a binary choice, we turn to models of binary choice that can be interpreted in a random utility framework. Such models include logit, probit, skewed logit, and complimentary log log (Greene 2012, p. 684; Nagler 1994). These models model the difference in utility between two possible actions as a linear combination of observed variables ($\boldsymbol{x\beta}$) plus a random variable (ε) that represents the difference in the unobserved random component of the utility of each option. Since ε is stochastic, these models can only predict a probability of choosing one action over the other.

Selecting the best model a priori is difficult because each has theoretical or practical advantages and disadvantages which we review in Section 3.8. However, for the coefficients of interest, all models come to essentially the same conclusions in terms of which coefficients are significant and the signs of those coefficients. All models also return similar predicted values over the range of interest. For the body of the paper we present the results from the probit model because it allows for easy comparison to the bivariate probit models necessary for some results.

We define the variable $LWBS_i$ to equal 1 if patient i abandons and 0 otherwise. We parametrize the basic probit model as follows

$$\begin{aligned}\text{Prob}\left(LWBS_i = 1|\mathbf{x}\right) =& \Phi(\beta_0 + \beta_1 OWAIT_i + \beta_2 CENSUS_i + \beta_3 OWAIT_i \times CENSUS_i \\ &+ \beta_4 TRITEST_i + \mathbf{X_i}\boldsymbol{\beta_P} + \mathbf{Z_i}\boldsymbol{\beta_T})\end{aligned} \tag{3.3}$$

where $\Phi(\cdot)$ represents the standard normal cumulative distribution function. $TRITEST_i$ is a binary variable indicating if any diagnostic tests were ordered for patient i at triage. $\mathbf{X_i}$ is a vector of patient-visit specific covariates including age, gender, insurance type, chief

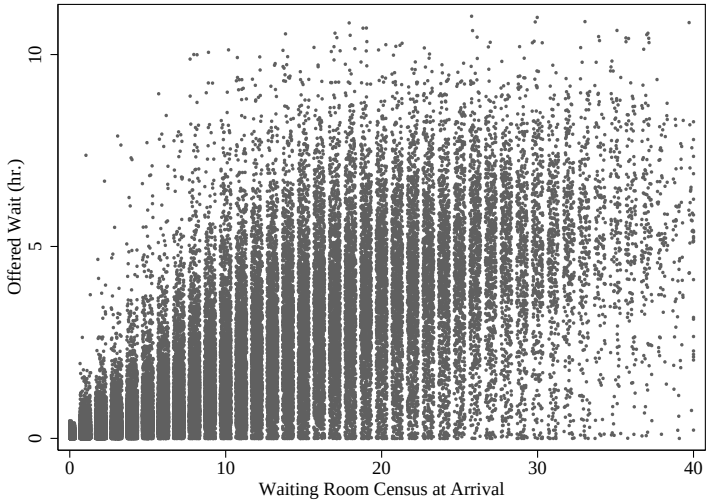
complaint, and pain level. $\boldsymbol{Z_i}$ is a vector of time related control variables including year, a weekend indicator, indicators for time of day by four-hour blocks, and the interaction of the weekend and time-of-day block variables. As we examine each of the hypotheses, we gradually add more variables to the model of Equation 3.3. We estimate the model separately for each triage level between 2 and 5.

The interaction term $OWAIT_i \times CENSUS_i$ is included to allow the marginal effect of $OWAIT$ to vary with $CENSUS$. If we were using ordinary least squares regression, a negative interaction coefficient would indicate that the marginal effect of $OWAIT$ is reduced when $CENSUS$ is high. However, due to the non-linear nature of the probit model, the interaction coefficient can not be interpreted in such a straightforward way. We discuss interpretation further in Subsection 3.7.2.1.

The $OWAIT$ variable is a bit different from all the other variables in the model in that it is not actually observed by the patient. Even for patients that enter service, the offered wait is not known until service begins, at which point abandoning is not an option. This variable should be thought of as an exposure variable. The offered wait is the maximum time a patient can spend in the system deciding whether to stay or abandon. The Erlang-A model is built around this idea that the longer a person is in the system, the higher her total probability of abandoning. Thus, the $OWAIT$ variable picks up this effect, that patients who are given the opportunity to be in the system longer are more likely to abandon, even though the actual offered wait value is not observed by the patient.

Our identification strategy is based on the assumption that $OWAIT$ and $CENSUS$ are not perfectly correlated and both contain exogenous variation. Essentially, we rely on the fact that treatment in the ED is a highly complex process with many "moving parts" (e.g., staffing levels, auxiliary services, coordination of many tasks and resources, etc.). This leads to high exogenous variation in treatment times for each patient, and this translates into high variance in offered wait times for waiting patients. This is seen in Figure 2 which shows the scatterplot of $OWAIT$ and $CENSUS$ (Waiting Room Census at Arrival) for ESI 3

Figure 2: Scatterplot of Offered Wait and Load for ESI 3 patients



Note: A small amount of circular noise or "jitter" has been added to help visualize the density of identical observations.

patients. Note that for any given level of $CENSUS$ there is a wide range of $OWAIT$.

A potential concern with this model specification is the collinearity between $OWAIT$ and $CENSUS$. The pairwise correlation between $OWAIT$ and $CENSUS$ is 0.72. However, the Variance Inflation Factors (VIF) for the model in Equation 3.3 range from 3.2 to 8.9 across triage levels, which is below the commonly accepted cutoff of 10 (Hair et al. 1995). Still, to be conservative, we mean center all stock and flow variables used in all models. When we do this for Equation 3.3, the VIFs range from 2.4 to 3.2, which is well within the acceptable range of collinearity.

When we examine Hypothesis 5, there is a potential endogeneity problem with the inclusion of the dummy variable indicating whether diagnostic tests were ordered at triage. The concern is that triage testing is not randomly assigned, but rather is assigned by a triage nurse based on the condition of the patient. As discussed in Batt and Terwiesch (2013), it is possible that there are unobserved variables, for example pallor, that are common to, or at least correlated with, both the triage test decision and the abandonment decision. For

example, a patient who arrives feeling terrible and looking terrible might be more likely to receive triage testing and less likely to abandon. This can bias not only the estimate of the coefficient of the triage test variable in the abandonment model, but can also bias all of the estimated coefficients.

We control for potential correlated omitted variables with a simultaneous equation model such as the bivariate probit model (Greene 2012). This model parametrizes both the triage test decision and the abandonment decision as simultaneous, latent-variable probit models as follows:

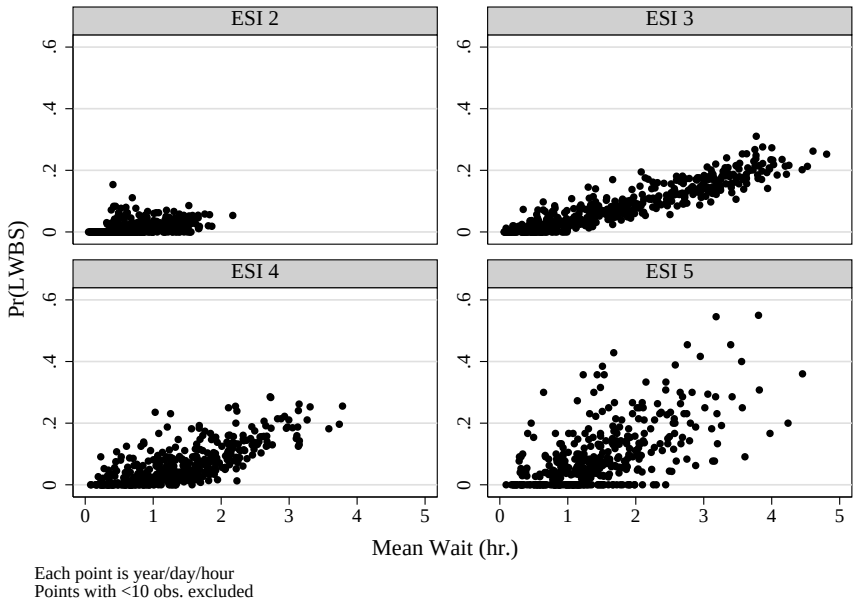
$$\begin{aligned} TRITEST_i^* =& \beta_{1,0} + \beta_{1,1} CENSUS_i + \beta_{1,2} BEDS_i + \mathbf{X_i}\boldsymbol{\beta_{1,P}} + \mathbf{Z_i}\boldsymbol{\beta_{1,T}} + \varepsilon_{1,i} \\ & TRITEST_i = 1 \text{ if } TRITEST_i^* > 0,\ 0 \text{ otherwise} \end{aligned} \tag{3.4}$$

$$\begin{aligned} LWBS_i^* =& \beta_{2,0} + \beta_{2,1} OWAIT_i + \beta_{2,2} CENSUS_i + \beta_{2,3} OWAIT_i \times CENSUS_i \\ & + \beta_{2,4} TRITEST_i + \mathbf{X_i}\boldsymbol{\beta_{2,P}} + \mathbf{Z_i}\boldsymbol{\beta_{T2,T}} + \varepsilon_{2,i} \\ & LWBS_i = 1 \text{ if } LWBS_i^* > 0,\ 0 \text{ otherwise} \end{aligned} \tag{3.5}$$

$\mathbf{X_i}$ and $\boldsymbol{Z_i}$ are specified as before in Equation 3.3. ε_1 and ε_2 are assumed to be standard bivariate normally distributed with correlation coefficient ρ. If $\rho = 0$, this indicates that the control variables are adequately controlling for the endogenous triage testing and the models can be estimated separately without significant bias.

Because approximately 60% of the patients in our data have multiple visits to the ED during the study period, we use the Huber/White/sandwich cluster-robust standard errors clustered on patient ID (Greene 2012). This adjusts the covariance matrix for the potential correlation in errors between multiple visits of a single individual. It also adjusts for potential misspecification of the functional form of the model. We find that this adjustment has very little effect on the results.

Figure 3: Pr(LWBS) vs. Wait Time



3.7. Results

3.7.1. Overview Graphs

Following the example of Zohar et al. (2002), we begin by using scatter plots to visualize the relationship between abandonment and wait time. If patients behave in accordance with the Erlang-A model such that wait time is the sole determinant of abandonment, then there should be a linear increasing relationship between expected wait time and probability of abandonment (Brandt and Brandt 2002, Zohar et al. 2002). Figure 3 shows the relationship of the probability of LWBS to the mean completed waiting time. Each dot represents a given year/day-of-week/hour-of-day combination. For example, one of the dots represents the mean wait and LWBS proportion of patients that arrived on Tuesdays of 2009 during the 4pm hour. Each graph has approximately 504 points (3 years $\times$ 7 days $\times$ 24 hours=504). However, points that represent less than 10 observations have been dropped. For example, there are not many ESI 5 patients at 4am on Mondays and that point has been dropped. Each subplot of Figure 3 is for a single triage or ESI level. In summary, each dot shows the

average wait time and percent of people who abandoned for patients that arrived at a given year/day/hour.

We observe several interesting features in Figure 3. First, there is a linear increasing trend for all triage levels (See Table 2 for the slope of a linear best-fit line.). While this is as expected, it is different from Zohar et al. (2002), in that Zohar et al. (2002) finds the surprising result that the probability of abandonment does not increase with expected wait (the linear fit is flat). This suggests that customers become *more* patient when the system is busy. We find no such evidence in the ED.

Table 2: Model Fit Measures of Regressing Pr(LWBS) on Wait Time

	Slope		RMSE	R^2
ESI 2	0.021	(0.002)	0.016	0.238
ESI 3	0.057	(0.001)	0.026	0.874
ESI 4	0.064	(0.003)	0.033	0.598
ESI 5	0.079	(0.005)	0.071	0.369

Secondly, the slope of the linear fit decreases with acuteness (Table 2). This suggests that sicker patients are less influenced by wait time, as one would expect.

The third feature we observe in Figure 3 is that the dispersion from the linear trend decreases with acuteness. Table 2 quantifies this effect by the root mean squared error (RMSE) for linear regressions for each of the graphs in Figure 3. Further, from the R^2 values in Table 2, we conclude that mean wait time is a very good predictor of abandonment probability for ESI 3. However, for ESI 4 and 5 patients, there appear to be other factors driving abandonment beyond just wait time. ESI 2 appears somewhat different. While ESI 2 displays a positive linear trend with little dispersion (significant positive slope and low RMSE), the model has the lowest R^2 further indicating that wait time explains very little of the the variation in ESI 2 abandonment probability. These differences in response across triage levels are particularly noteworthy when we recall that patients are not informed of their triage classification. Thus, the ESI triage system is doing a remarkable job of classifying people not only by medical acuity, but also by queuing behavior.

Given that wait time only partially explains the observed abandonment behavior, we now turn to patient-level regression models to better understand the operational drivers of abandonment.

3.7.2. Regression Analysis

The graphs in Section 3.7.1 are based on means calculated by aggregating across year/day/hour combinations. We now shift to patient-level analysis and use the binary-outcome probit regression models described in Section 3.6 to examine the hypotheses. Working at the patient level allows us to control for patient specific covariates such as age, gender, and insurance class, that we can not do as easily with the consolidated data in Section 3.7.1. For clarity, we focus on results for triage level ESI 3 in Subsections 3.7.2.1 and 3.7.2.2. We select ESI 3 because it has the largest number of observations, the highest abandonment rate, and the largest spread of wait times. We present comparisons across triage levels in Subsection 3.7.2.3, and in Subsection 3.7.2.4 we examine the impact of triage testing on all triage levels.

3.7.2.1 Observed Variables

Model 1 of Table 3 shows the results of estimating Equation 3.3 on the full sample. Probit coefficients are difficult to interpret directly since they represent a change in the linear z-score predictor due to a change in an independent variable. The first-order terms of Offered Wait and Census are positive and significant ($\beta_1, \beta_2 > 0$), but the negative interaction coefficient ($\beta_3 < 0$) makes it difficult to draw conclusions about hypotheses by inspection of the table. Estimated marginal effects and predicted values are more informative.

Because the model is nonlinear, the marginal effect of a covariate on the predicted probability is a function of not only the coefficients but also of the value of all the other covariates. To get a sense of the magnitude of effects, we calculate the mean marginal effect (across patients) of both the offered wait and census variables at their respective median values of 1.3 hours

Table 3: Effect of Wait Time, Census, and Flow on Pr(LWBS) [ESI 3]

	(1)	(2)	(3)	(4)
Offered Wait (hr.)	0.20***	0.12***	0.11***	0.11***
	(0.00)	(0.01)	(0.01)	(0.01)
Census	0.07***	0.06***	0.06***	0.06***
	(0.00)	(0.00)	(0.00)	(0.00)
Wait x Census	-0.01***	-0.01***	-0.01***	-0.01***
	(0.00)	(0.00)	(0.00)	(0.00)
Arrivals			0.01***	0.01***
			(0.00)	(0.00)
Depart(all)			-0.03***	
			(0.00)	
Depart(nonjump)				-0.03***
				(0.00)
Depart(jump)				-0.01
				(0.01)
N	65,622	35,855	35,855	35,855
BIC	32,767	28,780	28,721	28,729

Cluster robust standard errors in parentheses

Controls not shown: Age, Gender, Insurance, Pain, Triage Test, Year, Weekend×Block of Day

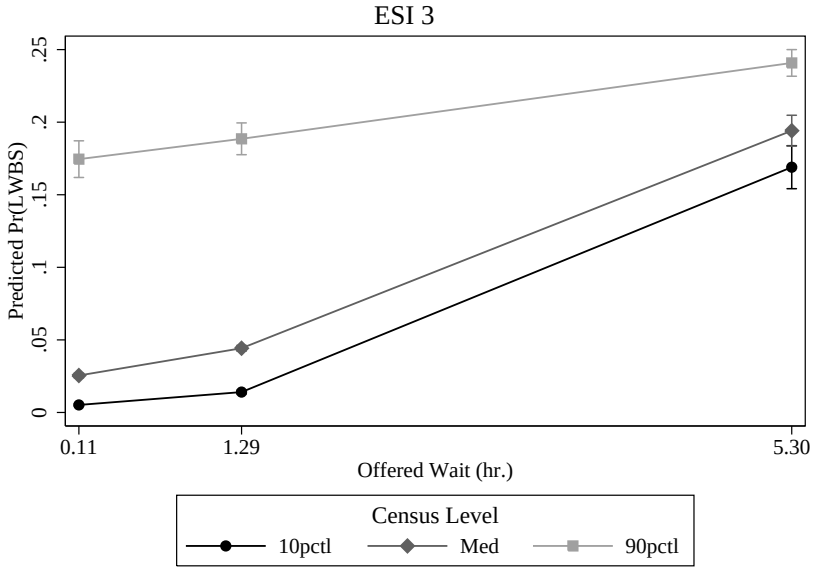
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

and 10 people. In Model 1, the predicted probability of abandonment increases by 2.0 percentage points with a one hour increase in offered wait. The marginal effect of observing an additional person in the waiting room when a patient arrives is a 0.5 percentage point increase in abandonment for ESI 3 patients. We can alternatively describe the marginal impact of an additional person in the waiting room as being equivalent to a 15 minute increase in offered wait. This supports Hypothesis 1 and shows that the Erlang-A model alone does not fully explain abandonment behavior. If it did, census should have no effect, controlling for wait time.

The marginal effect of waiting room census ranges from 0.1 to 0.4 percentage points for the other triage levels. Lu et al. (2012) estimates that a five person increase in queue length leads to a three percentage point drop in deli purchase incidence. This is equivalent to a marginal effect of 0.6 percentage points per person in line, and is quite close to our estimated marginal effect of 0.5 percentage points per person in the ED queue. This similarity in magnitude is somewhat surprising since waiting at the ED for medical care and waiting at the deli for cold cuts serve very different purposes and presumably generate markedly different levels of utility for the patients/customers.

Figure 4 shows the predicted abandonment probabilities at three levels of offered wait and census. Offered Wait is on the x-axis and the three test points (0.11, 1.29, 5.30 hours) are the 10th 50th, and 90th percentiles for ESI 3 patients. Each line on the graph represents the predicted probability of abandonment for a given census level. The three lines are the 10th, 50th, and 90th percentile census levels (1, 10, and 25 people respectively). The error bars represent the 95% confidence interval for the prediction. The upward slope of all of the lines conforms to the standard theory that longer waits lead to increased probability of abandonment. The vertical separation of the lines, however, indicates that patients are responding to the census level as well as the wait time. For example, a patient that arrives when the waiting room is relatively empty and experiences a 1.29 hour wait has a predicted probability of abandonment of 2%. However, if the waiting room is relatively crowded

Figure 4: Predicted Pr(LWBS) as a function of Offered Wait and Census



and all other covariates are held constant, the same patient has a predicted probability of abandonment of 19%. Thus, Figure 4 shows that patients respond to both increasing offered wait and waiting room census with increased abandonment.

The large gap between the median and 90th percentile census levels even for very short waits suggests that large crowds lead to rapid abandonment even when the actual wait time is low. This also explains why the slope of the 90th percentile census line is relatively flatter. People are likely abandoning sooner and are not remaining in the system to be impacted by the experienced wait. In other words, the impact of wait time is lower when the census is high. In contrast, for low to mid census levels, the effect of long wait times is larger.

To examine Hypothesis 2A, Hypothesis 2B, and Hypothesis 2C, we now include flow variables in the analysis. Recall that to do so we restrict the sample to those patients with an offered wait of greater than one hour, which reduces the sample size by almost half. Model 2 of Table 3 is the same as Model 1 (Equation 3.3) but with the restricted sample. We include it merely for comparison.

Model 3 of Table 3 adds in variables for the number of arrivals to the ED and for the number of departures into service. The positive and significant coefficient on arrivals supports Hypothesis 2A that arrivals lead to more abandonments. The coefficient on departures is significant and negative. This supports Hypothesis 2B that observing departures leads to reduced abandonment, presumably because waiting patients view these departures as a good sign of processing speed and progress towards service.

Model 4 of Table 3 splits the departures variable into nonjump and jump departures. The coefficient on nonjump departures is significant and negative while the coefficient on jump departures is insignificant. This continues to support Hypothesis 2B and suggests that Hypothesis 2C is correct. The insignificant effect of jump departures shows that any positive information about system speed is negated by the fact that the patient is getting jumped and is not moving closer to the head of the line. A one-sided z-test comparing the nonjump and jump coefficients confirms Hypothesis 2C and shows that the jump departures coefficient is larger (less negative) at a 94% confidence level. In terms of marginal effects, observing an arrival increases abandonment by 0.3 percentage points and observing a nonjump departure reduces abandonment by 0.6 percentage points. Figure 5 shows these same marginal effects in wait time equivalents. For example, observing an additional arrival per hour leads to the same increase in abandonment as an additional nine minutes of offered wait time. Similarly, observing a nonjump departure has the same impact on abandonment as a 19 minute reduction in offered wait.

In summary, patients respond to what they observe and the magnitudes of their responses are similar in magnitude to 10 to 20 minutes of waiting time.

3.7.2.2 Inferred Variables

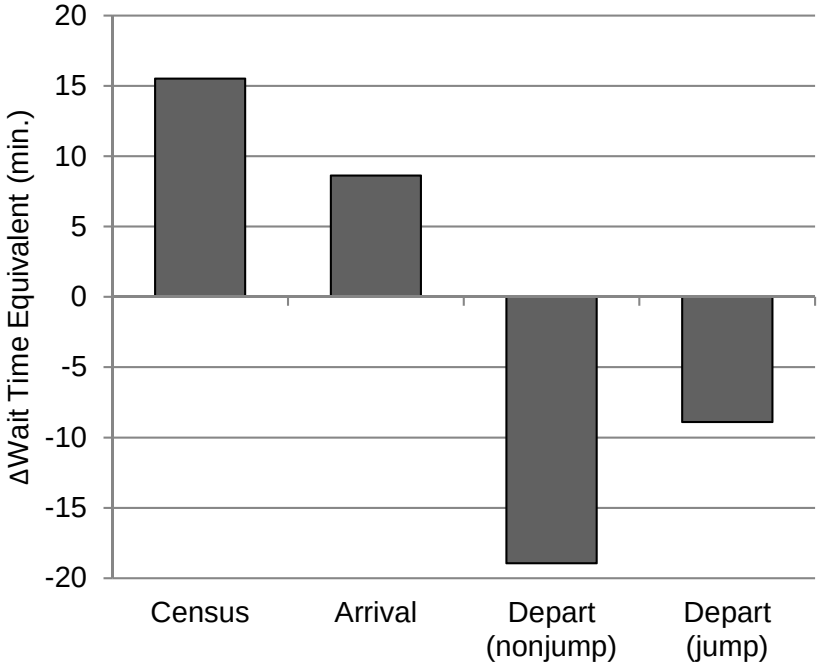
We now consider inferred system state variables. We are looking for evidence of patients behaving differently in the presence of patients that are ahead of or behind themselves in the priority queue structure. In practice, patients are not given any information about their own

Table 4: Effect of Wait Time and Census on Pr(LWBS) [Probit, ESI 3]

	(1)	(2)
Offered Wait (hr.)	0.19***	0.11***
	(0.00)	(0.01)
Census(Ahead)	0.09***	0.08***
	(0.00)	(0.00)
Census(Behind)	0.02***	0.01*
	(0.00)	(0.01)
WaitxCensus(Ahead)	-0.01***	-0.01***
	(0.00)	(0.00)
WaitxCensus(Behind)	-0.00***	-0.00
	(0.00)	(0.00)
Arrivals(Ahead)		0.05***
		(0.01)
Arrivals(Behind)		0.00
		(0.00)
Depart(Nonjump-Ahead)		-0.03***
		(0.00)
Depart(Nonjump-Behind)		-0.01*
		(0.01)
Depart(Jump-Ahead)		-0.06***
		(0.02)
Depart(Jump-Behind)		-0.01
		(0.01)
N	65,622	35,855
BIC	32,626	28,611

Cluster robust standard errors in parentheses
Controls not shown: Age, Gender, Insurance, Pain, Triage Test, Year, Weekend, Block of Day
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 5: Magnitude of Marginal Effect in Equivalent Minutes of Offered Wait



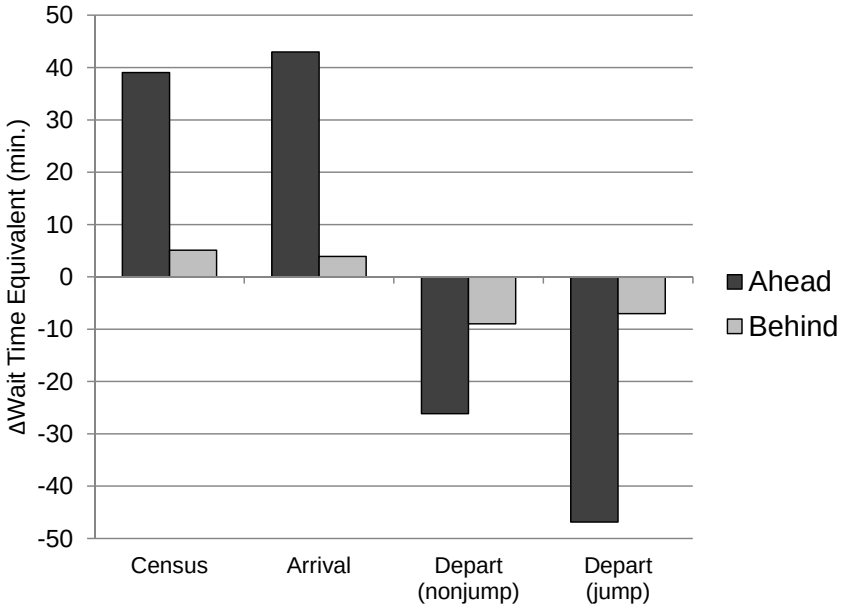
Note: Depart(jump) marginal effect estimate is statistically insignificant

priority level or other patients' priority levels. If patients truly have no information about the priority of those around them then one would expect the ahead and behind components of each queue status variable to have indistinguishable coefficients.

Model 1 in Table 4 is analogous to Model 1 in Table 3 but with the census variable split into ahead and behind components as described in Section 3.5. It is estimated on the full sample. A one-sided z-test shows that the Census(Ahead) coefficient is larger than the Census(Behind) coefficient. A Wald test of the marginal effects of Census(Ahead) and Census(Behind) confirms that patients respond more strongly to an increase in the census ahead than an increase in the census behind. This is all evidence in support of Hypothesis 3. The BIC of Model 1 in Table 4 is smaller than the BIC of Model 1 in Table 3 indicating that splitting the census into its ahead/behind components improves the fit of the model.

Model 2 in Table 4 is analogous to Model 4 in Table 3 but with the census and flow variables split into their respective ahead and behind components. We compare the coefficients of

Figure 6: Magnitude of Marginal Effects in Equivalent Minutes of Offered Wait



Note:None of the "Behind" estimates are statistically significant.

each ahead/behind pair and find that the values are significantly different and that the ahead component always has a larger magnitude than the behind component. This supports Hypothesis 4A, Hypothesis 4B, and Hypothesis 4C. Lastly, Model 2 in Table 4 has a smaller BIC than Model 4 in Table 3 indicating a better model fit with the stock and flow variables split into ahead/behind components.

Like Figure 5, Figure 6 shows the marginal effects of the split stock and flow variables in terms of equivalent wait time minutes. The marginal effect of the ahead component of each variable is much larger than of the behind component, and the magnitude of the effects on this subsample is much larger than for the full sample. We note that while the point estimates of the Depart(nonjump)Ahead and Depart(jump)Ahead seem quite disparate (-25 minutes and -45 minutes), they are statistically indistinguishable at the 10% level.

These results show that waiting patients respond quite differently to the presence and movement of patients of relatively higher and lower priority. The observed behavior is consistent with the idea that patients anticipate that it is largely the patients ahead of them in the

queue that interfere with their experience. While the directions of the effects are all as expected, this result is noteworthy because it shows that patients are indeed inferring relative priority information by observing the other patients.

We create a proxy measure of patients' classification accuracy by constructing the ratio

$$\theta = \frac{\beta_{AHEAD}}{\beta_{AHEAD} + \beta_{BEHIND}} \tag{3.6}$$

Let β_{AHEAD} be the estimated coefficient of one of the Ahead variables in Table 4 and let β_{BEHIND} be the estimated coefficient of the matching Behind variable. If patients believe that those behind them in line have no impact on residual wait time and if patients were perfect at classifying those ahead and behind, then β_{AHEAD}would be non-zero, β_{BEHIND} would be zero and θ would be unity. If, however, patients had no ability to discern those ahead and behind, then β_{AHEAD} would equal β_{BEHIND} and θ would equal 0.5 indicating that a patient's ability to classify other patients was no better than a coin toss. For example, if we focus on Jump Departures in Model 2, $\beta_{AHEAD} = -0.06$, $\beta_{BEHIND} = -0.01$, Looking at the other Ahead/Behind variable pairs in Table 4, we see θ range between 0.75 and 1. While we do not interpret θ as a literal measure of classification accuracy, it does suggest that patients are doing a fairly good job at classifying the other patients and responding accordingly.

3.7.2.3 Results Across Triage Levels

Table 5 shows the results of the best fitting model (Model 2 from Table 4) for all triage levels. The results are similar across triage levels in terms of which coefficients are significant and the signs of those coefficients. At first glance, there appear to be two unexpected results for ESI 4 (Model 3). The Census(Behind) coefficient is larger than the Census(Ahead) coefficient, and the Depart(Nonjump-Behind) coefficient is larger than the Depart(Nonjump-Ahead) coefficient. This would seem to suggest that ESI 4 patients are somehow more sensitive to

Table 5: Effect of Ahead/Behind variables on Pr(LWBS)

	(1)	(2)	(3)	(4)
	ESI 2	ESI 3	ESI 4	ESI 5
Offered Wait	0.14***	0.11***	0.15***	0.00
	(0.05)	(0.01)	(0.02)	(0.03)
Census(Ahead)	0.15***	0.08***	0.04***	0.04***
	(0.02)	(0.00)	(0.00)	(0.01)
Census(Behind)	0.02**	0.01*	0.08***	
	(0.01)	(0.01)	(0.02)	
WaitxCensus(Ahead)	-0.02***	-0.01***	-0.00***	-0.00
	(0.01)	(0.00)	(0.00)	(0.00)
WaitxCensus(Behind)	-0.00	-0.00	-0.01	
	(0.00)	(0.00)	(0.01)	
Arrival(Ahead)	0.03	0.05***	0.02***	0.02**
	(0.17)	(0.01)	(0.01)	(0.01)
Arrival(Behind)	0.01	0.00	0.01	-0.01
	(0.01)	(0.00)	(0.01)	(0.03)
Depart(Nonjump-Ahead)	-0.08***	-0.03***	-0.03***	-0.03***
	(0.02)	(0.00)	(0.01)	(0.01)
Depart(Nonjump-Behind)	-0.01	-0.01*	-0.05*	
	(0.01)	(0.01)	(0.03)	
Depart(Jump-Ahead)	0.07	-0.06***	-0.00	-0.01
	(0.22)	(0.02)	(0.02)	(0.02)
Depart(Jump-Behind)	-0.06	-0.01	-0.08	0.02
	(0.04)	(0.01)	(0.06)	(0.19)
N	8,974	35,855	19,745	5,213
BIC	2,688	28,611	9,568	3,593

Cluster robust standard errors in parentheses

Controls not shown: Age, Gender, Insurance, Pain,

Year, Weekend, Block of Day

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

those behind than in front of them. However a Wald test for coefficient equality shows that the two census coefficient are not significantly different at the 10% level, nor are the two depart coefficients. Thus, the correct interpretation is that ESI 4 patients do not appear to differentiate between those ahead of and behind in line, at least with regard to census level and departures.

ESI 5 is the most dissimilar of the four models. First, the variables $CENSUS\_BEHIND$ and $NONJUMP\_BEHIND$ are not included in the ESI 5 model because ESI 5 is the lowest priority level. Second, the Offered Wait has an insignificant effect on abandonment while Census(Ahead) continues to lead to greater abandonment. Without additional data on actual abandonment times, we are unable to determine if this result is because ESI 5 patients are truly insensitive to waiting time, or because they abandon so rapidly that the offered wait is irrelevant. Either way, it appears that for ESI 5 patients there is not much value in improving the wait time.

3.7.2.4 Triage Testing

Models 1 through 4 of Table 6 show the results of estimating the basic probit model of Equation 3.3 for ESI levels 2 through 5. In these models, the Triage Test coefficient is negative and significant indicating that those who receive an early diagnostic test order from the triage nurse are less likely to abandon. However, as described in Section 3.6 there is an endogeneity concern since triage testing is not randomly assigned. Models 5 through 8 of Table 6 show the results of estimating Equation 3.5 using a bivariate probit model. For ESI 3 and ESI 4 patients, the estimated correlation coefficient (ρ) is negative and significant indicating correlation in the error terms of Equations 3.4 and 3.5. This means that ESI 3 and 4 patients who receive triage testing are inherently more likely to stay. However, even after controlling for the correlation, triage testing continues to have a significant, albeit diminished, impact on abandonment, thus supporting Hypothesis 5. This confirms similar results reported in Pham et al. (2009). Once the correlation is controlled for, the marginal

Table 6: Effect of Triage Testing on Pr(LWBS) (Probit & Bivariate Probit models)

	Probit				Biprobit			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	ESI 2	ESI 3	ESI 4	ESI 5	ESI 2	ESI 3	ESI 4	ESI 5
Offered Wait	0.21***	0.20***	0.22***	0.11***	0.21***	0.19***	0.22***	0.11***
	(0.02)	(0.00)	(0.01)	(0.02)	(0.02)	(0.00)	(0.01)	(0.02)
Census	0.04***	0.07***	0.04***	0.05***	0.04***	0.07***	0.04***	0.05***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Wait x Census	-0.01***	-0.01***	-0.01***	-0.01***	-0.01***	-0.01***	-0.01***	-0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Triage Test (Y/N)	-0.44***	-0.51***	-0.47***	-0.23**	-0.41***	-0.14***	-0.25***	-0.46***
	(0.05)	(0.02)	(0.04)	(0.11)	(0.14)	(0.05)	(0.08)	(0.18)
ρ					-0.02	-0.23 ***	-0.15 ***	0.14
					(0.09)	(0.03)	(0.05)	(0.10)
Marginal Effects								
$\frac{\partial Pr(LWBS)}{\partial TRITEST}$	-0.014***	-0.064***	-0.031***	-0.024***	-0.013***	-0.018***	-0.018***	-0.043***
	0.002	0.002	0.002	0.010	0.004	0.007	0.005	0.012
N	27,455	65,622	39,806	10,483	27,455	65,622	39,806	10,483

Standard errors in parentheses

Controls not shown: Age, Gender, Insurance, Pain, Chief Complaint, Year, Weekend, Block of Day

Coefficients for Triage Testing equation not shown

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

effect of triage testing on abandonment is quite similar across ESI levels 2, 3 and 4, ranging from -1.3 percentage points to -1.8 percentage points.

The results for ESI 5 patients are slightly different in that the estimated correlation coefficient is positive, albeit insignificant (p-value: 0.18). This leads to the estimated coefficient on triage testing being larger in magnitude in the bivariate probit model than in the probit model. For ESI 5 patients, triage testing leads to a 4.3 percentage point reduction in abandonment probability, more than double the magnitude of the effect for the other triage levels. This suggests that the behavior of ESI 5 patients is more malleable than is the behavior of the more acute patients.

Failing to control for an endogenous regressor like triage testing has the potential to bias all coefficient estimates in the model. However, Table 6 shows that in our analysis, this does not appear to be a problem. The coefficients of the key variables of interest, offered wait and census, remain largely unchanged whether the probit or bivariate probit model is used. We perform the same bivariate probit analysis (not shown) on the best fitting model for all triage levels, similar to Table 5, and likewise find that while there is evidence of endogenous triage testing, controlling for it does not alter the estimates of the stock and flow variable coefficients. Thus we conclude that for the purpose of examining the effects of wait, census, and flows on abandonment, the simpler single equation model is sufficient.

3.8. Robustness of Model Selection

As mentioned in Section 3.6, there are several binary outcome models to choose from: logit, probit, skewed logit, and complimentary log log. These models differ in the choice of distribution of ε which determines the functional form of the response of the prediction to a change in an independent variable. Choosing either the logistic or the normal distribution leads to the well known logit and probit models, respectively. Assuming ε follows a complementary log log distribution ($F(\boldsymbol{x\beta}) = 1 - \exp[-\exp(\boldsymbol{x\beta})]$) leads to the CLL model. The Burr-10 distribution (Burr 1942) assumes ε is distributed with cumulative distribution

function $F(\boldsymbol{x\beta}, \alpha) = 1 - 1/\{1 + \exp(\boldsymbol{x\beta})\}^{\alpha}$. As a regression model, it is referred to as the skewed logistic or scobit model (Nagler 1994). Note that the logit model is a special case of the scobit model with $\alpha = 1$.

The logit and probit models are the most commonly used binary models and are quite similar, especially in the middle of the probability range. The logit has the further advantage of coefficients that can be immediately interpreted as impacts on odds-ratios. One advantage of the probit model is that it can be easily adapted to control for an endogenous regressor if necessary.

However, the logit and probit models are symmetric about $\boldsymbol{x\beta} = 0$, which imposes the restriction that observations with predicted probabilities close to 0.5 are most impacted by a change in the linear predictor. Since abandonment is a rare event (less than 10% of arrivals result in abandonment), the asymmetric cloglog and scobit models likely provide a better fit. Unlike the logit and probit models, the asymmetric models have a different fit depending on whether staying or abandoning is coded as "success." Thus we have at least six models to consider: logit, probit, CLL coded two ways, and scobit coded two ways.

Table 7 compares six such model specifications for the baseline model with offered wait, census, and the interaction for ESI 3 (cross-reference Table 3, Model 1). The top panel of the table shows estimated coefficients for the variables of interest. The middle panel shows marginal effects of the variables of interest at their respective medians. The bottom panel gives model fit statistics. We see that all the models are similar in terms of fit as indicated by both the log-likelihood and the BIC. The scobit (LWBS=1) model provides the best fit.

Comparing coefficient estimates across models is of limited use since the models are parametrized differently. However, we do see that all coefficients are significant and the signs are all in agreement. Further, comparing coefficients of the two versions of the cloglog model and the scobit model we see that the coefficients are dramatically different depending on whether stay or LWBS is coded as "success." This indicates that the data is skewed to one side, as

Table 7: Comparing Binary Response Models [ESI 3]

	(1) Logit	(2) Probit	(3) CLL (LWBS=1)	(4) CLL (Stay=1)	(5) Scobit (LWBS=1)	(6) Scobit (Stay=1)
Coefficients						
Offered Wait (hr.)	0.37***	0.20***	0.32***	-0.16***	0.63***	-0.17***
	(0.01)	(0.00)	(0.01)	(0.00)	(0.04)	(0.01)
Census	0.14***	0.07***	0.12***	-0.05***	0.21***	-0.06***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
Wait x Census	-0.02***	-0.01***	-0.02***	0.01***	-0.03***	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
alpha					0.12	11.2
					(0.015)	(5.43)
Marginal Effects						
Offered Wait	0.017***	0.020***	0.016***	-0.023***	0.023***	-0.022***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Census	0.005***	0.005***	0.004***	-0.006***	0.006***	-0.006***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N	65,622	65,622	65,622	65,622	65,622	65,622
log-likelihood	-16,262	-16,201	-16,314	-16,183	-16,177	-16,181
BIC	32,890	32,767	32,995	32,733	32,731	32,739

Cluster robust standard errors in parentheses

Controls not shown: Age, Gender, Insurance, Pain, Year, Weekend, Block of Day

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

expected.

Comparing marginal effects, we see again that the models all give similar results. A one hour increase in offered wait leads to a two to three percentage point increase in abandonment, or alternatively, a ten minute increase in offered wait leads to a 0.3 to 0.4 percentage point increase. A one unit increase in census leads to a 0.4 to 0.6 percentage point increase in abandonment. Note that the probit model, which we use for the presentation of main results in Section 3.7.2, underestimates the marginal effect of offered wait and census relative to the better fitting models. Thus, the results presented are conservative.

3.9. Discussion & Future Work

This study contributes to the understanding of customer waiting behavior by examining the queue abandonment behavior of patients waiting for treatment at a hospital emergency department. The essence of our contribution is in providing evidence that waiting customers glean information from watching the queue around them and update their utility function in response. Ours is among the first works to show customers responding to the actual functioning of the queue. We expand on prior work showing that the queue length (waiting room census, in our study) impacts behavior separate from wait time. This shows that in queues that are at least partially visible, the Erlang-A model does not fully capture abandonment behavior. Beyond just the queue length, we find that patients respond to other visual aspects of the queue in very sophisticated ways. For example, patients increase abandonment in response to observing arrivals, presumably because waiting patients recognize that the queue is not FCFS and the new arrivals may be served first. Further, waiting patients infer the relative priority status of those around them and respond differently to those more sick and less sick. For example, we find that the arrival of sicker, higher priority patients increases abandonment of those already waiting more so than does the arrival of less sick, lower priority patients. Waiting patients likely recognize that it is the sicker patients that will generally be served first. Lastly, we show that patients who have diagnostic tests ordered during triage are less likely to abandon. All of these effects are consistent

with patients updating their expected residual wait time in response to what they observe and experience. This is managerially relevant for any organization that wants to manage customer abandonment.

Throughout this work, we have intentionally avoided making any assumptions about the "optimal" level of abandonment. To do otherwise would require defining the hospital's objective function, but the hospital's objective is not at all clear. Revenue maximization would suggest eliminating abandonment and serving everyone who walks in the door. Likewise, a belief in a social obligation to serve all comers leads to a desire to eliminate abandonment. Social welfare maximization would suggest providing full information if the hospital believes that patients can accurately evaluate their own utility. However, if the hospital believes that patients are boundedly rational or can not accurately assess their need for treatment, then the hospital may withhold information. Lastly, profit maximization would suggest selectively serving only the most profitable patients while somehow avoiding serving the less profitable ones.

In our study hospital, the expressed objective is to minimize abandonment, largely out of a sense of duty to serve anyone seeking care. This is also a reasonable objective because the Centers of Medicare and Medicaid Services will soon require hospitals to report ED performance measures such as median wait time, median length of stay, and LWBS 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, hospitals will be looking to reduce abandonment at least to the target levels.

If we take minimization of abandonment to be the goal, then the managerial implication of our results is that the status quo of providing no information to the patients may not be optimal. Patient abandonment increased substantially with queue length, regardless of wait time, and thus either hiding the queue or providing more queue information may serve to reduce abandonment. The hospital could hide the queue by providing separate waiting rooms for each triage level, or it could provide more information in the form of a wait

time estimate or a queue status display board. Another implication of our results is that early initiation of service tends to reduce abandonment. Thus, the hospital could be more aggressive in ordering tests, perhaps even placebo tests, at triage.

Future work should use these findings to motivate and inform a series of controlled experiments. For example, it would be interesting to compare the effectiveness of providing more queue information versus obscuring information. Presumably, obscuring the queue would shift the behavior toward that of an invisible queue, such as a call center, but this should be explored empirically. Lessons learned from such experiments will serve to improve both ED management and our general understanding of human queuing behavior.

CHAPTER 4 : Doctors Under Load: An Empirical Study of State-Dependent Service Times in Emergency Care[1]

4.1. Introduction

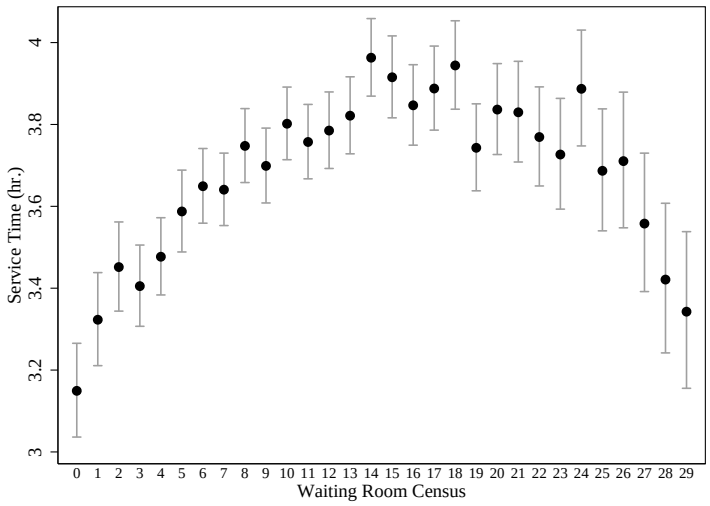
The Operations Management community has long been concerned with how crowding affects the performance of queuing systems. Basic queuing theory shows that crowding and high utilization of queues lead to exponentially increasing wait times. Since long waits are generally undesirable, it seems reasonable 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 load conditions.

In contrast, in domains such as transportation and telecommunications, high load 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 resources and/or servers that are not independent. For example, a highway lane is a shared resource for all the cars traveling in it and its performance can also be impacted by the traffic in adjacent lanes. Likewise, each node in a telecom network is a shared resource for many users, and it can be impacted by spillover from other nearby nodes (Gerla and Kleinrock 1980).

We bring these two 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 several reasons. First, the service (medical care) is provided by humans and as such is worker paced. Further, the required work for each patient is largely determined by the server (nurse or doctor) and the patient has limited

[1] This chapter is based on Batt, Robert J., Christian Terwiesch. 2013 "Doctors Under Load: An Empirical Study of State-Dependent Service Times in Emergency Care." Working Paper.

Figure 7: Service Time as a Function of Census



Notes: Mean and 95% confidence interval of mean shown. ED patients between 3pm and 11pm (second shift). Census is measured at the time a patient enters a treatment room.

knowledge of his or her own needs. This creates an environment in which the servers have a great deal of discretion over the encounter. This freedom can be used to alter both the service content (the specific tasks performed for the patient) and the service time (the total time to complete all tasks). Lastly, the ED is interesting because it is a complex service environment with many shared resources (nurses, doctors, equipment, hallways, laboratory, etc.). This suggests that the ED is prone to Slowdown.

Figure 7 previews our data, and motivates our study of Speedup and Slowdown mechanisms. The figure plots the mean service time of ED patients that arrive during second shift (3pm to 11pm) as a function of the waiting room census. Here, and throughout the paper, we define service time to be the time from when a patient is placed in a treatment bed to when treatment in the ED is complete as indicated by the patient either departing to go home or an inpatient bed request is placed in preparation for admission to the hospital. Thus service time does not include any time spent in the waiting room. The figure shows that mean service time rises from about 3.2 hours to 3.9 hours and then falls to 3.3 hours as the waiting room census ranges from low to high. If Speedup and Slowdown effects are monotone

in census level, then the non-monotone form of Figure 7 suggests that both Speedup and Slowdown are at work in the ED.

Prior empirical work on state-dependent service times has largely focused on the presence of state-dependent service times but not the mechanisms generating the state dependencies. In this paper, we identify and test for several state-dependent mechanisms including task reduction, early task initiation, multitasking, and interference. The first two are Speedup mechanisms and the latter two are Slowdown mechanisms.

Our study hospital has the additional feature of an "express lane" or FastTrack (FT) for low-acuity patients that is open only certain hours of the week. The FT is partially isolated from the rest of the ED operations; it uses dedicated treatment rooms and care providers. However, it relies on the same auxiliary services, such as the pathology lab and x-ray machines, as the main ED. We compare the effects of crowding on the ED and the FT.

We conduct a detailed econometric analysis of the service times and service content during more than 100,000 emergency department visits at a major U.S. hospital. We observe patient-level characteristics (age, gender, race, etc.) as well as timestamps of the progress of each visit including patient location and all laboratory, radiology, and medication orders. Survival analysis techniques are used to estimate the effects of Slowdown on service time and several common tasks. Count model regression techniques are used to identify various forms of service Speedup. Lastly, we use discrete event simulation to determine if these state-dependencies have a meaningful impact on the system. This research design allows us to make the following four contributions:

1. We examine several common ED tasks and find evidence of Slowdown in all. For example, time to first order (a measure of doctor speed) and medication delivery time time (a measure of nurse speed) increase by 26% and 11% respectively under high load.

2. We test for two Speedup mechanisms: early task initiation and task reduction. We

find strong evidence of early task initiation with the expected number of triage tests increasing from 0.3 to 0.9 in the ED. We find only limited use of task reduction in the ED, while task reduction is more common in the FT.

3. We show that the net effect of Speedup and Slowdown is different in the ED and the FT. In the ED, service time first increases then decreases with load as the relative strength of Speedup and Slowdown mechanisms shifts. In the FT, Speedup and Slowdown balance out leading to little change in service time with increased crowding.

4. We show that models which ignore the state-dependent service times overestimate the system utilization and congestion.

These findings offer several operational insights for managers. 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. This suggests that care providers should consider incorporating state-dependencies into ED care protocols. For both the healthcare domain and other domains, our findings show that understanding the micro-level mechanisms behind state-dependent service rates is important for properly modeling service systems where the server has discretion over the service speed and the service content. Our results, particularly regarding task reduction and task time increases, suggest an operational explanation for the many studies that have shown a link between crowding and reduced clinical quality in the ED (e.g., Fee et al. 2007, Pines and Hollander 2008). However, in this paper we remain focused on the effect of crowding on service time.

4.2. Clinical Setting

Our study is based on data from a large, urban, teaching hospital with an average of 4,700 ED visits per month over the study period of January, 2009 through December, 2011. The ED has 25 treatment rooms and 15 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. 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 operates somewhat autonomously from the rest of the ED in that it utilizes seven dedicated beds and is usually staffed by dedicated group of Certified Registered Nurse Practitioners (CRNP) rather than Medical Doctors (MD)[2] .

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.

The study hospital operates in a manner similar to many hospitals across the United States. Upon arrival, 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 official 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. 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).

After triage, all patients wait in a common waiting room to be taken to a treatment room. 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 serve triage level 4 and 5 patients

[2]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.

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. For example, the ED might serve a triage level 4 patient if the patient has been waiting a long time and there are not more acute patients that need immediate attention. Similarly, the FT might serve a triage level 3 patient if the patient has been waiting a long time and the patient's needs can be met by the nurse practitioners in the FT. 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 marks the beginning of the service time. Soon after being moved to a treatment room, a physician meets with and examines the patient.[4] 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, medications, or both are made at this point. All lab test, radiology scan, and medication orders are recorded electronically in the patient tracking system, but 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 scanner and then back to the treatment room.

Eventually, the physician decides that either the patient can leave or the patient needs to

[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.

be admitted. If the patient is to be admitted, a bed request is entered in the inpatient bed management system. At this point, ED service is considered complete. The patient waits for an available inpatient bed and is considered a "boarder" in the ED. This boarding period can be quite long with a mean of 3.6 hours. During this time, the patient continues to occupy a treatment 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, service time ends when the patient leaves the ED. Mean service time for admitted and discharged patients is 3.6 hours and 3.8 hours respectively.

For patients served by the FT, the care process is quite similar to that in the ED, except with a dedicated group of rooms and providers. Once in a treatment room, the care provider evaluates the patient, orders any necessary tests and medicines, and attempts to provide treatment as rapidly as possible. Just as in the ED, all lab test, radiology scan, and medication orders are recorded electronically in the patient tracking system. One difference between the FT and the ED is that there is a less clear demarcation between provider and nurse tasks. For example, a CRNP treating a FT patient may order and deliver medications him or herself, whereas in the ED, the doctor would order the medicine and the nurse would deliver it. However, as in the ED, FT labs are generally drawn by a nurse and scan orders enter the same transport queue as the ED patients. When treatment is complete, the patient is discharged. In rare cases, the FT provider can reroute the patient to the ED or admit the patient to the hospital. Mean service time for FT patients is 1.3 hours.

4.3. Framework & Hypotheses

We are interested in examining the mechanisms of state-dependent service times at the server level. We begin with the assumption from classical queuing theory that the service time distribution is not affected by the system state (Wolff 1989). However, as seen in Figure 7, it appears that this assumption is false in our setting, and that there is a dependence between the system state and the service time. Similarly, Armony et al. (2012) includes

an empirical examination of an ED at the system level and finds evidence of both Speedup and Slowdown. However, in contrast to what we show in Figure 7, Armony et al. (2012) finds that the ED first speeds up and then slows down as load increases from low to high. Armony et al. (2012) muses (but does not test) that Speedup may be the result of rushing as care providers respond to a mild increase in congestion, and that Slowdown could also be caused by factors such as fatigue, shared resources being spread thin, or nurses having to devote too much time to caring for boarding patients.

We posit that there are several mechanisms that may be at work and that these can be classified by the direction of their impact on service times and by the number of resources involved. In the following we describe these mechanisms, their related prior research, and the hypotheses they motivate.

4.3.1. Slowdown

We focus first on Slowdown, or mechanisms that increase service time. Prior literature has shown that both fatigue and multitasking can lead to Slowdown in individual servers. 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.

In our setting, multitasking refers to a single resource, such as a nurse, being simultaneously responsible for multiple patients, but individual tasks are not necessarily performed simultaneously. For example, a nurse may deliver a medication to one patient and then draw blood from a second patient. In effect, the nurse acts as a single channel server performing tasks for different patients in rapid succession. As the nurse becomes responsible for more patients and gets "spread thin," the arrival rate of tasks to the nurse's virtual queue increases leading to longer completion times for each individual task from the patient's point of view. The Psychology literature on human multitasking shows that multitasking additionally incurs

cognitive switching costs which further hinder productivity (Pashler 1994). These switching costs increase with increased levels of multitasking. See Kc (2012) for a summary of this literature. Kc (2012) empirically examines the effect of ED physician multitasking on service time and finds that multitasking leads to longer service times. A shared resource, like an x-ray machine, can be thought of as multitasking in a similar manner. With more patients in treatment, more x-ray requests are generated, the queue for x-rays grows, and the completion time for each x-ray increases.

Another form of Slowdown can occur with multiple resources. As mentioned in Section 4.1, the idea of high load causing Slowdown is well established in fields such as transportation and telecommunications (Chen et al. 2001, Gerla and Kleinrock 1980). In these settings, this effect is commonly referred to as congestion. However, we refer to this as *interference* since this is a different effect than is generally referred to in the Operations Management literature by the word "congestion." In the Operations Management literature, congestion usually refers to long queues and long wait and sojourn times, but does not imply any change in service times. In the transportation and telecommunications settings, and in this paper, the Slowdown effect of interest is an increase in the actual service time, regardless of wait time. In the ED, examples of interference are crowded hallways that slow workers down and nurses waiting for computer terminals.

Both multitasking and interference are conceptually similar to queuing models with shared processors (e.g., Yamazaki and Sakasegawa 1987, Aksin and Harker 2001). Shared processor models assume that the server (or servers) splits its processing capacity across all items in service leading 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 rel-

evant 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. Similarly, Jaeker and Tucker (2012) report evidence of interference caused by high load levels in a hospital leading to longer inpatient stays.

To test for Slowdown, it is not sufficient to simply examine total service time for a patient because the service time is affected by both Speedup and Slowdown effects. To isolate and test for the existence of Slowdown, we focus on the durations of a few specific tasks that are common to many ED visits such as lab specimen collection time and x-ray completion time. We suspect that such tasks are susceptible to all the Slowdown mechanisms described above. For example, lab collection time will increase as a nurse juggles more patients, becomes fatigued, and has to wait in line to use the pneumatic tube system to send a sample to the lab. Thus, while we do not attempt to separately identify the Slowdown mechanisms at work, we test for the presence of Slowdown in general, and we expect crowding to lead to increased task times.

Hypothesis 1 Task time increases with load: $\frac{\partial TaskTime}{\partial Load} > 0$

4.3.2. Speedup

Turning now to Speedup, or mechanisms that decrease service times, 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 reduce the service time.

Perhaps the simplest form of service time reduction is *rushing*. That is, the server simply works faster. 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 Speedup under load with restaurant waiters and loan application processors, respectively.

Since rushing affects task time, we are actually testing the net effect of Slowdown and rushing when we test for the effect of load on task time in Hypothesis 1. We have stated Hypothesis 1 as we have ($\frac{\partial TaskTime}{\partial Load} > 0$) because we believe that Slowdown dominates rushing in the ED. In fact, we believe that rushing is not prevalent in many knowledge-intensive services such as the ED. Despite what is portrayed on TV, doctors and nurses are rarely seen running through the halls of the ED or performing specific procedures faster.

4.3.2.1 Task Reduction

Papers by Hopp et al. (2007) and by Alizamir et al. (2011) build on the optimal queue control stream and suggest another Speedup mechanism; *task reduction*. Hopp et al. (2007) describes a service system with discretionary task completion that is 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 policy that sets a service cutoff time for every value of queue length. This policy is monotone decreasing in queue length. Alizamir et al. (2011) 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 (i.e. task completions per time interval), 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.

The discretionary task completion model of Hopp et al. (2007) forms the basis of our hypotheses regarding task reduction. In the Hopp et al. (2007) framework, the variable under the server's control is service time itself. In our setting, we assume the variable under the physician's control is the service content, that is the quantity of diagnostic tests ordered. Further, we assume that utility is concave increasing with the number of tests. As long as reducing testing quantity reduces service time, the insight from Hopp et al. (2007) that service time should be reduced under crowding translates to the hypothesis that testing should be reduced under crowding. This leads to the following two hypotheses.

Hypothesis 2 Service time increases with diagnostic testing: $\frac{\partial ServiceTime}{\partial Tests} > 0$

Hypothesis 3 Diagnostic testing decreases with load: $\frac{\partial Tests}{\partial Load} < 0$

The idea that service time should be reduced under crowding seems quite reasonable, perhaps even obvious, in the settings proposed in Hopp et al. (2007) such as telemarketers and salespeople. However, in a medical setting such as an ED, the idea of reducing the quantity -and perhaps quality- of care for Mrs. Jones just because she has the bad luck of being in the ED when there is a crowd seems less obvious. We leave that discussion for later and simply draw on the Hopp et al. (2007) model to suggest an interesting hypothesis, that physicians change the thoroughness of their testing based on crowding. We refer to this behavior with the admittedly loaded term "cutting corners."

4.3.2.2 Early Task Initiation

While rushing and task reduction are Speedup mechanisms that can be implemented by a single server, we propose the mechanism of *early task initiation* as a Speedup mechanism that may exist between resources. Early task initiation is similar to concurrent engineering, which for nearly thirty years has been acknowledged as an effective way to speed up product

development cycles. First widely publicized by Imai et al. (1985) and Takeuchi and Nonaka (1986), the concept is to take logically consecutive tasks and execute them with some amount of temporal overlap. This requires the decision makers at each task to make some guesses or bets since the exact needs of the other tasks are not yet known. The fundamental tradeoff is that overlapping the tasks reduces the time to market but that too much overlap leads to rework or poor final design quality (Loch and Terwiesch 1998).

A similar opportunity exists in multi-resource service systems. A service task may be started early, before it is even fully known if the task is required. For example, in the ED, as described in Section 4.2, triage nurses have the option of ordering some diagnostic tests.[5] If tests are ordered at triage, the tests can be processed while the patient is waiting in the waiting room. Then when the patient sees the physician the tests are already under way or may even be ready for review. This reduces service time. However, the downside of triage testing is that the nurse is "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. This over-testing is undesirable because it increases financial costs, medical risk for the patient (if the test is risky), and load on the diagnostic resources.

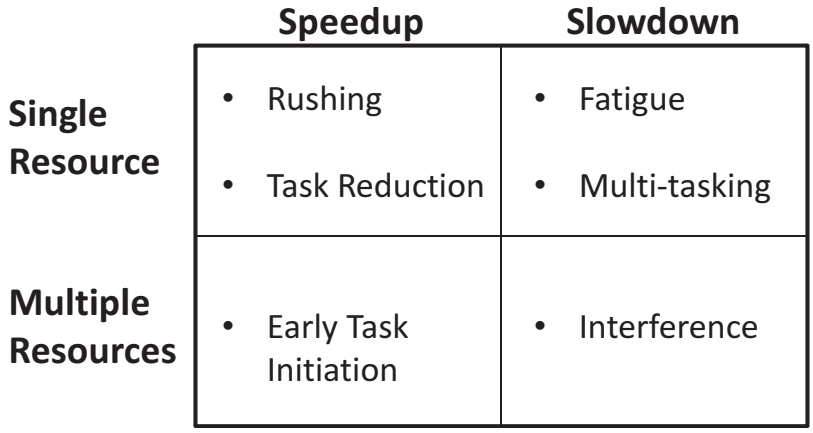
Note that the benefits of ordering tests at triage are largest when waiting times are long. This is because much or all of the test processing time occurs in parallel with the patient waiting in the waiting room. Conversely, when waiting times are short, there is little benefit to triage testing since the service time will be reduced by only a few minutes. However, the consequences of over-testing do not scale with load in a similar fashion, and therefore we hypothesize that triage testing will be most common when the system is crowded.

Hypothesis 4 Triage testing increases with load: $\frac{\partial TriageTest}{\partial Load} > 0$

For early task initiation to be beneficial, an increase in triage testing should lead to a

[5] These triage tests are commonly referred to as Advanced Triage Protocols in the medical community.

Figure 8: State-Dependent Mechanisms



decrease in doctor testing. If triage nurses have perfect information we would expect a one for one trade-off between triage and doctor testing; each incremental triage test would lead to a one test reduction in doctor testing. However, if the nurses have imperfect information and "betting" is an apt description, then we would expect the marginal triage test to lead to a reduction in doctor testing of less than one.

Hypothesis 5 Doctor testing decreases less than one unit for each unit increase in triage testing: $-1 < \frac{\partial DocTest}{\partial TriageTest} < 0$

4.3.3. Net Impact on Service Time

Figure 8 summarizes the categorization of the mechanisms just described that potentially lead to state-dependent service times. Since Speedup and Slowdown mechanisms work in opposing directions, the net impact is indeterminate a priori. Therefore, we do not posit an hypothesis. Nonetheless, it is worth examining the net change in service time with load to determine the relative magnitudes of the two effects. Based on Figure 7, we suspect that Slowdown dominates but that Speedup effects eventually become large enough such that the marginal effect of load is negative. Stated differently, we believe that for low to mid level loads $\frac{\partial ServiceTime}{\partial Load} > 0$, and for mid to high level loads $\frac{\partial ServiceTime}{\partial Load} < 0$.

4.3.4. Additional Related Literature

While we have already referenced the prior work to which our study is most closely related, we also point out connections to two other bodies of literature.

Our work is influenced by the portion of the analytical queuing theory literature has been stimulated by problems in the health care domain. Topics such as capacity planning (e.g., Lee and Zenios 2009, Allon et al. 2011), staffing (e.g., deVericourt and Jennings 2011, Yankovic and Green 2012) and patient flow (e.g., Green et al. 2006a, Ibrahim and Whitt 2011) have all been studied extensively. We direct the reader to Green (2006) for an overview of this literature. This body of work has largely been focused on characterizing and managing service systems from a high-level or system design point of view.

Our work also relates to the large body of medical literature on crowding's effect on service and quality. Many of these papers 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) report a positive relationship between crowding and length of stay while Lucas et al. (2009) find no significant relationship. McCarthy et al. (2009) report that crowding drives up wait times but has no effect on service times, a result that agrees with traditional queuing theory.

Our contribution to the literature is in bringing attention to the level of the servers (care providers). We expand on the prior literature by providing detailed evidence of both Speedup and Slowdown mechanisms occurring simultaneously. By focusing at the micro-level, we can identify the underlying mechanisms that lead to the service time changing under load. We hope this will extend the understanding of service system productivity.

4.4. Data Description & Definitions

Our 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 8 provides descriptive statistics of the patient population. For much of the analysis,

Table 8: Summary Statistics of Patients

	ED	FT
Variable	Mean	Mean
Age	41.2 (0.05)	34.6 (0.08)
Female	61% (0.002)	59% (0.003)
Triage 2	25.1% (0.001)	1.3% (0.001)
Triage 3	59.2% (0.001)	5.3% (0.001)
Race: Black	58.6% (0.002)	64.3% (0.001)
Race: White	24.8% (0.001)	19.8% (0.002)
Diagnostics Ordered	5.38 (0.014)	1.27 (0.010)
Service Time (hr.)	3.77 (0.009)	1.31 (0.006)
N	108,014	36,427

Standard error in parentheses

we focus on a single chief complaint at a time since the testing patterns and response to crowding can be quite different from one chief complaint to another. Chief complaint is determined by the triage nurse, and our data contains 129 unique chief complaints. The two most common chief complaints in the ED are abdominal pain and chest pain, representing 13% and 9% of the ED visits respectively. The two most common chief complaints in the FT are limb pain and body pain, representing 14% and 9% respectively.

We are primarily concerned with how load affects ED performance. In the ED, there are several census measures that indicate system load. These include waiting room census, ED in-service census, FT in-service census, and ED boarding census. To calculate these census measures, we divide the study period (2009-2011) into 15-minute intervals labeled t, and we use the patient visit timestamps to generate the census variables $WAIT_t$, $EDSERV_t$, $FTSERV_t$, and $BOARD_t$ as the number of patients in the given location during interval t.

When we examine task times (Hypothesis 1), we perform the analysis at the per-hour level

and thus we generate the load variables $\overline{WAIT_h}$, $\overline{EDSERV_h}$, $\overline{FTSERV_h}$, and $\overline{BOARD_h}$ as the average for hour h for each of the census measures

For the rest of our analysis, we focus solely on the waiting room census as the measure of ED load. We do this because observation and anecdotal evidence suggests that ED nurses and doctors focus on this number as a key indicator of the crowd level in the ED. Further, the waiting room census is visible to the triage nurses and the rest of the ED staff on electronic dashboards. We also choose to focus on waiting room census because it effectively has no upper bound and thus has a great deal of variability. In contrast, in-service and boarding census measures are limited by the number of beds in the ED. Lastly, we focus on waiting room census because we believe that the effects of crowding in the ED primarily occur when the ED is operating in a highly-loaded or overloaded state with all treatment beds filled.

We assign two load measures to each patient visit: load at arrival, $aLOAD_i$, and load at the start of service, $sLOAD_i$. For example, for patient i who arrives at time interval $t = 1$ and is put in a treatment room at time $t = 8$, $aLOAD_i = WAIT_1$, and $sLOAD_i = WAIT_8$. We then convert the variables $aLOAD_i$ and $sLOAD_i$ into vectors of dummy variables $\widetilde{\mathbf{aLOAD}}_i$ and $\widetilde{\mathbf{sLOAD}}_i$ corresponding to low, mid, and high census levels. The cut points are set such that 25% of observations are in each of the low and high categories and 50% of the observations are in the mid category. For $\widetilde{\mathbf{aLOAD}}_i$, the cut points are at 5 and 19, while for $\widetilde{\mathbf{sLOAD}}_i$ the cutpoints are at 4 and 18.

One reason for using a categorical load variable is that it allows for a more general response to load than would including just linear and quadratic terms of $LOAD_i$. The other reason is that it greatly simplifies the reporting of results and comparison of various models as will be seen in Section 4.6.

We examine several dependent variables in this study including task time, service time, and the counts of various categories of diagnostic tests.

To study task timing, we define the variable $\overline{TASKTIME_h}$ as the mean task completion

time across all tasks of a given type ordered during hour h. The tasks we examine are as follows:

First Order Time: The time from when a patient is put in a treatment room until the first order (lab, scan, or medication) is recorded.

Lab Collection Time: The time from a lab order being placed until the nurse closes out the order indicating that the specimen has been sent for analysis.

Medication Delivery Time: The time from a medication order being placed until the nurse closes out the order indicating the medication has been given to the patient.

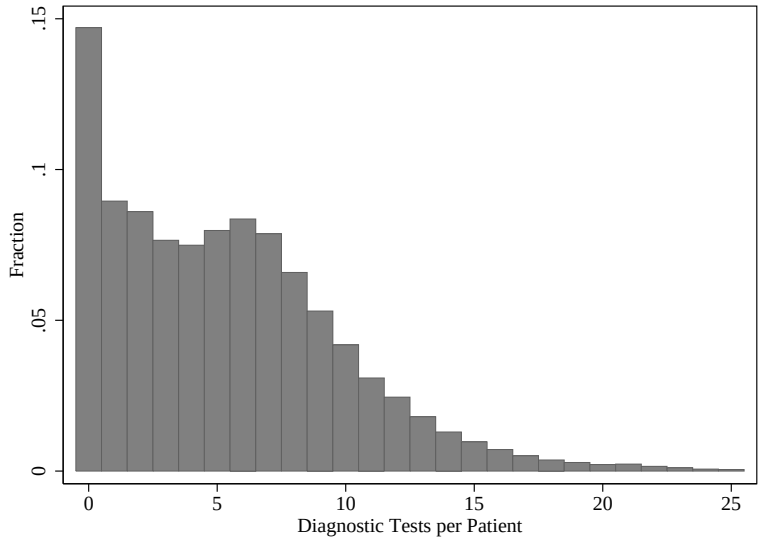
Scan Completion Time: The time from a radiology scan order being placed until the patient returns from having the scan performed. This does not include the time required for a radiologist to perform the official "reading" of the scan.

The first task is a proxy for the physician busyness level. The second and third tasks are proxies for nurse busyness. The fourth task measures the sojourn time for an auxiliary service that is shared by the entire ED and by other parts of the hospital, depending on the scan type.

The service time variable, $SERVTIME_i$, is defined as the time from placement in a treatment room until the patient is either discharged or a bed request is placed for admission to the hospital for patient i. Note that service time does not include any time spent in the waiting room.

The last major dependent variable is the count of diagnostic tests ordered either by the triage nurse or doctor. There are two types of diagnostic tests: lab tests and radiology imaging scans. Lab tests are chemical analyses of patient tissue or fluid such as urinalysis, white blood cell counts, and electrolyte levels. Most of these tests are performed by the hospital's central pathology lab that serves both the ED and the rest of the hospital. Radiology imaging scans include various types of electromagnetic and ultrasonic imaging techniques,

Figure 9: Number of Diagnostic Tests per ED Patient



such as x-ray, magnetic resonance imaging, and computed tomography, used to view the internal structures of the body. For most of our analyses we aggregate these two types of tests into a single variable $TEST_i$ (Figure 9). We also decompose diagnostic test orders into $TRITEST_i$ and $DOCTEST_i$ based on whether the test was ordered at triage or in the treatment room. The average ED patient receives 0.6 triage tests and 4.8 doctor tests, however 15% receive no diagnostic tests at all. The mean number of diagnostic tests varies significantly by chief complaint and triage level. For some models, we further decompose $TRITEST_i$ and $DOCTEST_i$ into the number of labs and scans ordered at each location.

$$TRITEST_i = TRILAB_i + TRISCAN_i \tag{4.1}$$

$$DOCTEST_i = DOCLAB_i + DOCSCAN_i \tag{4.2}$$

4.5. Econometric Specification

We now develop the econometric specifications for testing our hypotheses. In the discussion below, the index h indicates an hour in the study period, and the index i denotes a patient visit to the emergency department.

To test Hypothesis 1, we are interested in how load impacts the duration of various common ED tasks, thus we turn to survival analysis models. Specifically, we use an accelerated-failure-time (AFT) model with a log-normal distribution. The AFT model relates the log of service time to a vector of covariates and a random error term ϵ through a linear equation. For this analysis, we relate the mean task time in a given hour to a load variable and control variables as follows:

$$ln(\overline{TASKTIME_h}) = \alpha + \beta_1 \overline{WAIT_h} + \beta_2 \overline{EDSERV_h} + \beta_3 \overline{FTSERV_h} + \beta_4 \overline{BOARD_h} + \mathbf{Z}_i \boldsymbol{\phi} + \epsilon_h \tag{4.3}$$

$\boldsymbol{Z}_i$ is a vector of time related control variables including year, month, day of week, hour of day, and the interaction of day of week and hour of day. Because our dependent variables are estimated means, we use weighted least squares to estimate the model where the weights are equal to the number of tasks ordered in hour h (Wooldridge 2009). Also, because the data forms a time series with possible autocorrelation we use the Newey-West covariance estimator to provide standard errors that are robust to both heteroskedasticity and autocorrelation (Greene 2012). Due to these complications, we must assume that ϵ_h follows a normal distribution. Thus, Equation 4.3 is an AFT model with a log-normal underlying distribution. In this specification, positive coefficients β or $\boldsymbol{\phi}$ indicate an increase in mean task time, and Hypothesis 1 is supported if $\beta > 0$.

We note that the AFT model implies specific assumptions about the underlying survival and hazard functions. Specifically the log-normal specification implies a hazard function

that is first increasing and then decreasing. We choose this distribution because this form resembles the hazard function form of the data and because it allows us to correct for the weighting and autocorrelation as mentioned above. The major advantage of the AFT model over the semi-parametric Cox proportional hazard model is that the AFT model coefficients can be directly interpreted as changes in duration and a prediction of mean task time can be calculated.

Hypothesis 2 examines the effect of testing on service time. We achieve this by using the following AFT model specification which includes variables for both labs and scans ordered at triage and by the doctor.

$$\begin{aligned} ln(SERVTIME_i) =& \alpha + \widetilde{\mathbf{aLOAD}}_i\boldsymbol{\beta} + \delta_1 TRILAB_i + \delta_2 DOCLAB_i \\ &+ \delta_3 TRISCAN_i + \delta_4 DOCSCAN_i + \mathbf{W}_i\boldsymbol{\theta} + \mathbf{Z}_i\boldsymbol{\phi} + \epsilon_i \end{aligned} \tag{4.4}$$

The dependent variable is now service time for patient i. $\mathbf{W}_i$ is a vector of patient-visit specific covariates such age, gender, race, triage level, and chief complaint. $\boldsymbol{Z}_i$ is again a vector of time related control variables including year, month, hour of day and a weekend indicator variable. $\widetilde{\mathbf{aLOAD}}_i$ is a vector of dummy variables indicating mid and high load with the low load condition as the omitted category. We now assume ϵ follows a log-logistic distribution rather than a log-normal distribution. While the log-logistic and log-normal distributions assume similarly shaped hazard functions, we use the log-logistic function here because it better fits the data based on the Bayesian Information Criterion. Positive values of the δ coefficients support the hypothesis that testing leads to longer service times.

Hypotheses 3, 4, and 5 all require examining how test order quantities change with respect to some load or testing variable. Since the dependent variable is discrete and fairly small, we need to use a count-type model. Further, as seen in Figure 9, 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. The ZINB model combines a binary logit process with prob-

ability density $f_1(\cdot)$ and a negative binomial count process with probability density $f_2(\cdot)$ to create the combined density

$$f(y|\mathbf{x}) = \begin{cases} f_1(1|\mathbf{x}_1) + \{1 - f_1(1|\mathbf{x}_1)\} f_2(0|\mathbf{x}_2) & \text{if} \quad y = 0 \\ \{1 - f_1(1|\mathbf{x}_1)\} f_2(y|\mathbf{x}_2) & \text{if} \quad y \geq 1 \end{cases} \tag{4.5}$$

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 the negative binomial count process. This model has the conditional mean

$$E[y|\mathbf{x}] = \frac{1}{1 + \exp(\mathbf{x_1}\boldsymbol{\eta_1})} \times \exp(\mathbf{x_2}\boldsymbol{\eta_2}) \tag{4.6}$$

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. 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 $\boldsymbol{\eta}_2$.

To test for the presence of task reduction (Hypothesis 3) we examine how $DOCTEST_i$ changes with load controlling for $TRITEST_i$. We formulate the linear predictors $\mathbf{x}_{i,1}\boldsymbol{\eta}_1$ and $\mathbf{x}_{i,2}\boldsymbol{\eta}_2$ as follows:

$$\mathbf{x}_{i,j}\boldsymbol{\eta}_j = \alpha_j + \widetilde{\mathbf{sLOAD}}_i\boldsymbol{\beta}_j + \delta_j TRITEST_i + \mathbf{W}_{i,j}\boldsymbol{\theta}_j + \mathbf{Z}_{i,j}\boldsymbol{\phi}_j \text{ for } j = 1, 2 \tag{4.7}$$

Similar to Equation 4.4, $\mathbf{W}_{i,j}$ is a vector of patient-visit specific covariates such as age, gender, race, triage level, and chief complaint. $\mathbf{Z}_{i,j}$ is a vector of time related control

variables such as year, month, shift, and a weekend indicator variable.[6]

To test for the presence of early task initiation (Hypothesis 4), we switch to $TRITEST_i$ as the dependent variable of the ZINB model. We formulate the linear predictors as follows:

$$\mathbf{x}_{i,j}\boldsymbol{\eta}_j = \alpha_j + \mathbf{a\widetilde{LOAD}}_i\boldsymbol{\beta}_j + \mathbf{W}_{i,j}\boldsymbol{\theta}_j + \mathbf{Z}_{i,j}\boldsymbol{\phi}_j \text{ for for } j = 1, 2 \tag{4.8}$$

To test the marginal impact of triage testing on doctor testing (Hypothesis 5), we use the model specified in equation 4.7 but focus on the marginal effect of $TRITEST$ rather than of $\widetilde{sLOAD}$.

While we do not offer an hypothesis for the net impact of Speedup and Slowdown on service time, we are interested in the empirical result. Since we are again looking at a duration outcome, we use the following AFT model:

$$ln(SERVTIME_i) = \alpha + \mathbf{a\widetilde{LOAD}}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\theta} + \mathbf{Z}_i\boldsymbol{\phi} + \epsilon_i \tag{4.9}$$

This model is the same as equation 4.4 minus the lab and scan count variables. In this specification, positive coefficients $\boldsymbol{\beta}$, $\boldsymbol{\theta}$,or $\boldsymbol{\phi}$ indicate an increase in service time.

4.6. Results

To test for evidence of Slowdown effects, we examine the impact of load on task times (Hypothesis 1). Tables 9 and 10 show the results for the ED and the FT respectively.

The general pattern we see in both the ED and the FT is that task times increase as load increases, which supports Hypothesis 1. We also see that the in-service census for the given area (ED or FT) tends to be the main driver of the increase, which supports the

[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.

Table 9: Effect of Load on Task Times (ED only)

	(1)		(2)		(3)		(4)	
	1st Order Delay		Lab Collect Time		Med Time		Scan Time	
Wait Census	0.001	(0.001)	0.005***	(0.001)	0.002**	(0.001)	0.004***	(0.001)
ED In-Service	0.031***	(0.001)	0.006***	(0.001)	0.014***	(0.001)	0.013***	(0.002)
FT In-Service	-0.001	(0.003)	0.002	(0.003)	0.008**	(0.004)	0.019***	(0.005)
Boarding	0.007***	(0.001)	0.021***	(0.002)	0.012***	(0.002)	0.011***	(0.002)
N	24,465		21,278		25,344		25,424	

Newey-West HAC robust standard errors in parentheses

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

Table 10: Effect of Load on Task Times (FT only)

	(1)		(2)		(3)		(4)	
	1st Order Delay		Lab Collect Time		Med Time		Scan Time	
Wait Census	0.006***	(0.001)	0.002	(0.003)	0.004	(0.004)	0.003	(0.002)
ED In-Service	0.003	(0.002)	0.010*	(0.006)	0.004	(0.006)	-0.005	(0.004)
FT In-Service	0.092***	(0.007)	0.087***	(0.014)	0.052***	(0.017)	0.076***	(0.012)
Boarding	-0.000	(0.003)	0.025***	(0.008)	0.004	(0.007)	0.004	(0.005)
N	10,247		5,449		6,387		7,585	

Newey-West HAC robust standard errors in parentheses

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

idea of nurse or doctor multitasking leading to increased service times. To get a sense of the magnitude of change in task times, we note that the interquartile range of $EDSERV$ spans from 15.5 patients to 23 patients; a range of 7.5 patients. Multiplying 7.5 by the ED In-Service coefficient and exponentiating the product gives the percent change in the dependent variable. For example, the First Order Delay for ED patients increases by about 26% ($exp(7.5 \times 0.031) = 1.26$) as the number of patients in the ED service beds ranges from the 25th to 75th percentile. That other census measures are significant for some models and not others shows that Slowdown is caused by different factors for different tasks. Still, the general finding remains the same; task times increase with load.

For most of the rest of our analysis, the variable of interest is the three-level load variable. Because of this, we generally report predicted values and pairwise differences between predicted values. This provides a more intuitive interpretation than simply reporting regression coefficients, especially for the ZINB models with two coefficients for each variable. Also, for all models, we run and report the results separately for various subsets of the population. We show results for both the ED and the FT to allow for comparison between these two systems. Also, we show aggregate results for all chief complaints and then for each of the most common chief complaints in the ED and the FT individually. We do this because aggregating patients across chief complaints forces the coefficients of all the variables to be the same across all chief complaints. For example, in the aggregate model, the difference in testing between low and high crowding is the same regardless of whether the patient has a heart attack or a tooth ache. While this is perhaps tolerable for the load variable, it is outright dubious for other variables such as age and gender. By focusing on a single chief complaint at a time we sacrifice sample size but gain tenability.

As we turn our attention to task reduction (Hypothesis 3), we first show that diagnostic tests do indeed increase service time (Hypothesis 2). Table 11 shows the results of estimating Equation 4.4. All coefficients are positive or insignificant. The exponentiated form of these coefficients can be interpreted as multipliers of the service time. For example, for

Table 11: Effect of Diagnostic Orders on Service Time

	ED			FastTrack		
	(1)	(2)	(3)	(4)	(5)	(6)
	All ED	AP	CP	All FT	LP	BP
TRILAB	-0.001	0.018***	-0.002	0.023***	0.023	0.057***
	(0.002)	(0.004)	(0.006)	(0.007)	(0.043)	(0.020)
DOCLAB	0.024***	0.038***	0.019***	0.142***	0.121***	0.139***
	(0.001)	(0.002)	(0.002)	(0.003)	(0.009)	(0.010)
TRISCAN	0.015***	-0.025	0.108***	0.108***	0.086***	0.216***
	(0.005)	(0.036)	(0.015)	(0.007)	(0.011)	(0.033)
DOCSCAN	0.154***	0.175***	0.186***	0.371***	0.295***	0.514***
	(0.002)	(0.004)	(0.007)	(0.005)	(0.009)	(0.016)
Controls						
Age, Race, Gender	Yes	Yes	Yes	Yes	Yes	Yes
Chief Complaint	Yes	AP only	CP only	Yes	LP only	BP only
Triage	1-5	2,3	2,3	1-5	3-5	3-5
Doctor	Yes	Yes	Yes	No	No	No
Year, Month, Weekend, Hour	Yes	Yes	Yes	Yes	Yes	Yes
N	98,304	12,449	8,499	36,300	5,111	3,103

Standard errors in parentheses

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

AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

an abdominal pain patient, each doctor-ordered lab increases the service time by about 4% (exp(0.038) = 1.039). Also note that the doctor-ordered test coefficient is always significantly larger than the related triage-ordered test coefficient. This speaks to the time savings provided by early task initiation (Hypothesis 4), discussed below.

For task reduction, we examine how the quantity of doctor-ordered tests changes with load, controlling for tests ordered at triage (Table 12). For ED patients in aggregate, column 1

Table 12: Doctor Tests (controlling for triage testing)

	ED			FastTrack		
	(1)	(2)	(3)	(4)	(5)	(6)
	All ED	AP	CP	All FT	LP	BP
Predicted Doctor Orders						
Wait Census: Low	4.81	6.67	5.58	0.96	1.13	1.00
	(0.026)	(0.070)	(0.082)	(0.021)	(0.055)	(0.081)
Wait Census: Mid	4.75	6.63	5.49	0.89	1.04	1.03
	(0.016)	(0.049)	(0.052)	(0.011)	(0.031)	(0.037)
Wait Census: High	4.88	6.73	5.52	0.86	0.89	1.02
	(0.029)	(0.090)	(0.091)	(0.015)	(0.041)	(0.059)
Differences						
Mid vs Low	-0.064**	-0.039	-0.084	-0.065***	-0.088	0.029
	(0.030)	(0.088)	(0.099)	(0.023)	(0.063)	(0.089)
High vs Low	0.065	0.058	-0.059	-0.091***	-0.235***	0.019
	(0.042)	(0.122)	(0.132)	(0.027)	(0.071)	(0.105)
High vs Mid	0.129***	0.097	0.025	-0.026	-0.147***	-0.010
	(0.033)	(0.101)	(0.104)	(0.019)	(0.052)	(0.071)
Controls						
Age, Race, Gender	Yes	Yes	Yes	Yes	Yes	Yes
Triage	1-5	2, 3	2, 3	3-5	3-5	3-5
Triage Test Count	Yes	Yes	Yes	Yes	Yes	Yes
Doctor	Yes	Yes	Yes	Yes	Yes†	Yes
Year, Month, Weekend, Shift	Yes	Yes	Yes	Yes	Yes	Yes
N	98,583	12,482	8,517	35,751	5,113	3,103

Standard errors in parentheses

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

AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

†Variable included in count portion of model only

shows a small but statistically significant dip in testing at mid level crowding suggesting some amount of cutting corners. Columns 2 and 3 provide no evidence of cutting corners on specific chief complaints in the ED. The results look quite different for FT patients. Columns 4 and 5 show strong evidence of task reduction for FT patients in aggregate and for limb pain patients in isolation. For example, the predicted mean number of doctor ordered tests drops from 1.13 to 0.89 as load goes from low to high. There is no evidence of cutting corners with body pain patients (Column 6).

To test for early task initiation (Hypothesis 4), we examine how triage testing changes with load (Table 13). Note that in this table we do not separate by ED and FT since that distinction is not made until after triage when the patient is placed in a treatment bed. Thus, we show the results for all patients and for the four most common chief complaints. We see that across the board, triage testing increases with load. For example, the predicted mean number of triage tests for an abdominal pain patient almost triples from 0.397 to 1.019 and roughly quadruples from 0.342 to 1.309 for a chest pain patient as load goes from low to high. This is strong evidence in support of Hypothesis 4. We also examine how doctors and nurse practitioners respond to triage testing (Hypothesis 5). Table 14 shows the marginal effect $\frac{\partial DOCTEST}{\partial TRITEST}$ for several levels of $TRITEST$. Almost all of the marginal effects are between negative one and zero indicating that doctors are reducing testing in response to triage testing, but not at a one-for-one ratio. This supports the idea of there being uncertainty in the triage nurse ordering. Further, for ED patients (columns 1 and 2), the marginal effect of $TRITEST$ approaches zero for larger values of $TRITEST$ indicating decreasing marginal benefit of triage testing. This shows that when the triage nurse orders just one test, there is a high probability that this is a useful test and the doctor can reduce her testing orders by one. However, as more triage tests are ordered, the uncertainty in their usefulness increases and each additional test leads to smaller reductions in doctor testing. In contrast, the marginal benefit of triage testing is much smaller for FT patients. This shows that early task initiation is less effective in the FT.

Table 13: Count of Triage Tests

	(1) All (ED&FT)	(2) AP	(3) CP	(5) LP	(6) BP
Predicted Triage Orders					
Wait Census: Low	0.312	0.397	0.342	0.272	0.276
	(0.005)	(0.015)	(0.021)	(0.013)	(0.014)
Wait Census: Mid	0.547	0.822	0.843	0.470	0.477
	(0.004)	(0.015)	(0.020)	(0.012)	(0.012)
Wait Census: High	0.742	1.019	1.309	0.474	0.550
	(0.009)	(0.031)	(0.042)	(0.020)	(0.023)
Differences					
Mid vs Low	0.235***	0.424***	0.501***	0.197***	0.201***
	(0.007)	(0.022)	(0.0310)	(0.018)	(0.019)
High vs Low	0.430***	0.622***	0.967***	0.201***	0.274***
	(0.011)	(0.036)	(0.048)	(0.025)	(0.028)
High vs Mid	0.195***	0.198***	0.466***	0.004	0.073***
	(0.010)	(0.034)	(0.047)	(0.023)	(0.026)
Controls					
Age, Race, Gender	Yes	Yes	Yes	Yes	Yes
Triage	1-5	1-5	1-5	1-5	1-5
Chief Complaint	Yes	AP only	CP only	LP only	BP only
Year, Month, Weekend, Shift	Yes	Yes	Yes	Yes	Yes
Weekend×Shift	Yes	Yes	Yes	Yes	Yes
N	144,252	14,351	9,689	10,536	10,099

Standard error in parentheses

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

AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

Table 14: Marginal Effect of Triage Testing on Doctor Testing

	ED				FastTrack			
	(1)		(2)		(3)		(4)	
	AP		CP		LP		BP	
TRITEST								
0	-1.09	(0.05)	-0.99	(0.05)	-0.27	(0.06)	-0.11	(0.10)
1	-0.96	(0.04)	-0.88	(0.04)	-0.35	(0.04)	-0.14	(0.04)
2	-0.84	(0.03)	-0.79	(0.03)	-0.33	(0.02)	-0.15	(0.06)
3	-0.73	(0.02)	-0.70	(0.02)	-0.21	(0.02)	-0.14	(0.05)
4	-0.64	(0.01)	-0.62	(0.02)	-0.10	(0.02)	-0.12	(0.03)
N	12,482		8,517		5,113		3,103	

Standard error in parentheses

AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

Finally, we look at the net effect of crowding on service time. Table 15 shows the results of the log-logistic AFT regression of service time (Equation 4.9). Columns 1, 2, and 3 show the results for ED patients. We find evidence of service time first rising and then falling a bit as load moves from low to mid to high. This result matches the pattern seen in Figure 7. This suggests that Slowdown effects strongly dominate at first but then as load continues to increase Speedup effects increase and bring the service time back down. However, there is no evidence of Speedup ever being so strong as to reduce the high-load service times below the low-load service times. In contrast, in the FT, there is little evidence of load having any effect on service time. In Column 4 we see an increase of 0.03 hours (1.8 minutes) in service time for all FT patients when going from low to mid load, but no other predicted differences are significant. These results show that in the ED, Slowdown is the dominant result of crowding, while the FT is largely immune from crowding affecting service times.

4.7. 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: triage testing and patient abandonment. Triage 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 to the researcher (e.g., countenance, sweating,

Table 15: Mean Service Time Predictions and Differences

	ED			FastTrack		
	(1)	(2)	(3)	(4)	(5)	(6)
	All ED	AP	CP	All FT	LP	BP
Predicted Mean Service Time						
Wait Census: Low	3.97	5.22	3.68	1.42	1.74	1.48
	(0.02)	(0.06)	(0.06)	(0.02)	(0.04)	(0.06)
Wait Census: Mid	4.13	5.49	3.83	1.45	1.77	1.56
	(0.01)	(0.05)	(0.04)	(0.01)	(0.03)	(0.04)
Wait Census: High	4.04	5.45	3.69	1.46	1.70	1.56
	(0.02)	(0.09)	(0.07)	(0.01)	(0.04)	(0.05)
Differences						
Mid vs Low	0.156***	0.271***	0.150**	0.029*	0.030	0.079
	(0.022)	(0.078)	(0.071)	(0.017)	(0.050)	(.063)
High vs Low	0.063**	0.232**	0.011	0.034	-0.037	0.086
	(0.031)	(0.111)	(0.093)	(0.022)	(0.063)	(0.081)
High vs Mid	-0.093***	-0.038	-0.139*	0.005	-0.068	0.007
	(0.024)	(0.091)	(0.072)	(0.016)	(0.045)	(0.059)
Controls						
Age, Race, Gender	Yes	Yes	Yes	Yes	Yes	Yes
Chief Complaint	Yes	AP only	CP only	Yes	LP only	BP only
Triage	1-5	2,3	2,3	1-5	3-5	3-5
Doctor	Yes	Yes	Yes	No	No	No
Year, Month, Weekend, Hour	Yes	Yes	Yes	Yes	Yes	Yes
N	98,304	12,449	8,499	36,300	5,111	3,103

Standard error in parentheses

Stars displayed for differences only: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

pallor). However, triage testing influences the testing decision of the doctor (the coefficient on $TRITEST$ in Equation 4.7 is significant in all models), and thus it can be considered a "treatment." Just like the triage testing decision, the doctor testing decision 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 triage testing and is possibly driven by the same unobservable covariates affecting triage testing and doctor testing. Thus, there is the potential for a three-way interaction between triage testing, abandonment, and doctor testing. For example, a patient with chest pain who is pale and sweaty may have an increased probability of receiving diagnostic tests both in triage and from the doctor, and might be highly likely to wait to be served since the patient feels quite sick. This would lead to positive uncontrolled correlations among the three equations. Note, however, that all these potential issues only become problematic if the observed covariates are not rich enough to capture the differences between patients. Also, if there is a bias, it is likely that the bias is toward sicker patients remaining and being tested during high crowds. This would be a bias against our hypotheses, and thus our findings are conservative.

The "ideal" test for endogeneity would be a three-equation model that simultaneously estimates the endogenous treatment (triage testing), the self-selection (abandonment), the resulting zero-inflated count outcome (doctor testing) and the respective 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. Overall, 6.5% of patients abandon the queue. However, the rate ranges from 3% under low crowding to 12% under high crowding. 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:

$$\begin{gathered} S^* = \alpha_1 + \widetilde{\mathbf{aLOAD}}\beta_{\mathbf{1}} + \delta_1 \mathbf{1}(TRITEST > 0) + \mathbf{W_1}\boldsymbol{\theta_1} + \mathbf{Z_1}\boldsymbol{\phi_1} + \varepsilon_1 \\ STAY = 1 \text{ if } S^* > 0,\ 0 \text{ otherwise} \end{gathered} \tag{4.10}$$

$$\begin{gathered} D^* = \alpha_2 + \widetilde{\mathbf{sLOAD}}\beta_{\mathbf{2}} + \delta_2 TRITEST + \gamma_2 FT + \mathbf{W_2}\boldsymbol{\theta_2} + \mathbf{Z_2}\boldsymbol{\phi_2} + \varepsilon_2 \\ DOCTEST\_YN = 1 \text{ if } D^* > 0,\ 0 \text{ otherwise} \end{gathered} \tag{4.11}$$

The vectors $\mathbf{W_1}$and $\mathbf{W}_2$contain the patient covariates age, gender, race, chief complaint, and triage level. The variable FT is a dummy variable indicating if the patient was treated in the FastTrack. The vector $\mathbf{Z_1}$ contains controls for year, month, weekend, and shift, while the vector $\mathbf{Z_2}$ contains controls for only weekend and shift. We drop the year and month variables from the second equation to provide an exclusion restriction to help with model identification even though the model technically is identified by the non-linearity of the probit equations. ε_1 and ε_2 are assumed to be standard bivariate normally distributed with correlation coefficient ρ, and Equation 4.11 is only observed if $STAY = 1$. 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. We see in Table 16 that indeed the estimated correlations are insignificantly different from zero for models 2, 3, and 5, but for models 1 and 4, the correlation is positive and significant. The coefficients in the upper panel show that the probability of staying (not abandoning) decreases with load, as one

Table 16: Heckman Probit Selection model of Abandonment and Doctor Testing

	(1) All	(2) AP	(3) CP	(4) LP	(5) BP
Stay (Y/N)					
Wait Census Mid	-0.469***	-0.744***	-0.625***	-0.210***	-0.576***
	(0.018)	(0.053)	(0.075)	(0.065)	(0.070)
Wait Census High	-0.890***	-1.411***	-1.174***	-0.538***	-0.995***
	(0.020)	(0.059)	(0.088)	(0.076)	(0.080)
Doctor Test (Y/N)					
Wait Census Mid	-0.043***	-0.018	-0.204***	-0.069**	0.006
	(0.011)	(0.045)	(0.077)	(0.034)	(0.039)
Wait Census High	-0.052***	-0.023	-0.361**	-0.175***	0.072
	(0.016)	(0.073)	(0.148)	(0.044)	(0.058)
ρ	0.148***	-0.201	0.657	0.636***	-0.302
	(0.043)	(0.213)	(0.294)	(0.119)	(0.233)
Age, Race, Gender, Triage	Yes	Yes	Yes	Yes	Yes
Chief Complaint	Yes	AP only	CP only	LP Only	BP Only
FastTrack†	Yes	Yes	Yes	Yes	Yes
Year††, Month††, Weekend, Shift	Yes	Yes	Yes	Yes	Yes
N	144,252	14,351	9,689	10,536	10,099

Standard errors in parentheses

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

†Variable included in Doctor Test Y/N portion of model only

††Variable included in selection (Stay Y/N) portion of model only

AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

would expect. The coefficients in the lower panel indicate that for all patients in aggregate and for chest pain and limb pain patients (columns 1, 3, and 4), doctors are less likely to order tests during high crowding, whereas in Table 12 we only saw limited evidence of cutting corners under load. These results show that while the observed covariates are controlling for much of the patient differences, correcting for the remaining correlation between self-selected abandonment and doctor testing only strengthens our findings.

We also check for unobserved correlation between triage testing and patient abandonment. We use a bivariate probit model similar to the selection model above, but without needing to adjust for the selected sample. Table 17 shows that the census coefficients are all significant and in the direction we expect; crowding increases triage testing and abandonment. We also see that models 1, 2, and 5 show significant positive correlation in the errors. However, if we repeat the analysis for patients of a single triage level at a time, then the correlation becomes insignificant. Together, these two sets of results suggest that patient abandonment may create a bias in the results, but any bias that does exist makes our findings conservative since the correlations are all positice. Further, these robustness checks suggest that the bias can largely be corrected for with our control variables and by focusing on a single triage level at a time.

To examine the potential endogeneity between triage testing and doctor testing we again use a bivariate probit model. We ignore the middle step of abandonment based on the above results showing that there is not a significant bias. The results of this analysis are mixed in that some models show significant between-equation correlation, and others do not (Table 18). The coefficients in the upper panel are all as expected indicating increased triage testing with increased crowding. With the exception of Column 6, the coefficients in the lower panel are as expected, showing either no change or a decrease in doctor testing with load, controlling for triage testing. Column 6 shows a slight increase in doctor testing when crowding is at the mid level. However, the two load dummy variables (Wait Census Mid & Wait Census High) are jointly insignificant and the fit of the model actually improves if the

Table 17: Bivariate Probit of Triage Test and Stay/LWBS

	(1) All ED	(2) AP	(3) CP	(4) LP	(5) BP
Triage Test (Y/N)					
Wait Census Mid	0.496***	0.631***	0.685***	0.426***	0.434***
	(0.011)	(0.031)	(0.037)	(0.038)	(0.039)
Wait Census High	0.646***	0.726***	0.942***	0.447***	0.457***
	(0.014)	(0.039)	(0.046)	(0.048)	(0.049)
Stay (Y/N)					
Wait Census Mid	-0.413***	-0.677***	-0.658***	-0.273***	-0.505***
	(0.019)	(0.059)	(0.084)	(0.081)	(0.070)
Wait Census High	-0.804***	-1.326***	-1.221***	-0.588***	-0.920***
	(0.023)	(0.070)	(0.103)	(0.081)	(0.081)
ρ	0.264***	0.184**	-0.078	-0.422	0.412***
	(0.037)	(0.084)	(0.125)	(0.237)	(0.123)
Age, Race†, Gender	Yes	Yes	Yes	Yes	Yes
Chief Complaint	Yes	AP only	CP only	LP Only	BP Only
Triage	1-5	1-5	1-5	1-5	1-5
Year, Month, Weekend, Shift	Yes	Yes	Yes	Yes	Yes††
N	107,825	13,802	9,193	10,536	10,099

Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

†Race included in Triage Test portion of model only

††Month included in Triage Test portion of model only

AP: Abdominal Pain, CP: Chest Pain, LP: Limb Pain, BP: Body Pain

Table 18: Bivariate Probit of Triage Testing and Doctor Testing

	ED			FastTrack		
	(1)	(2)	(3)	(4)	(5)	(6)
	All ED	Abd. Pain	Chest Pain	All FT	Limb Pain	Body Pain
Triage Test (Y/N)						
Wait Census Mid(a)	0.571***	0.672***	0.713***	0.322***	0.347***	0.303***
	(0.013)	(0.033)	(0.039)	(0.024)	(0.052)	(0.075)
Wait Census High(a)	0.819***	0.891***	1.062***	0.385***	0.407***	0.343***
	(0.016)	(0.044)	(0.049)	(0.029)	(0.060)	(0.085)
Doctor Test (Y/N)						
Wait Census Mid(s)	-0.037***	-0.037	-0.117**	-0.024	-0.087*	0.118*
	(0.013)	(0.046)	(0.057)	(0.020)	(0.051)	(0.062)
Wait Census High(s)	-0.031*	-0.047	-0.211***	-0.038	-0.214***	0.067
	(0.018)	(0.062)	(0.073)	(0.024)	(0.058)	(0.069)
ρ	-0.060 ***	-0.022	-0.172***	-0.136***	-0.409***	0.008
	(0.010)	(0.0932)	(0.034)	(0.019)	(0.041)	(0.068)
Age	Yes	Yes	Yes	Yes	Yes†	Yes†
Race	Yes	Yes	Yes	Yes†	Yes†	Yes†
Gender	Yes	Yes	Yes	Yes	Yes	Yes
Chief Complaint	Yes	AP only	CP only	Yes	LP only	BP only
Triage	1-5	2,3	2,3	3-5	3-5	3-5
Year	Yes	Yes	Yes	Yes	Yes	Yes†
Month	Yes†	Yes†	Yes†	Yes†	Yes†	Yes†
Weekend	Yes	Yes	Yes	Yes	Yes†	No
Shift	Yes	Yes	Yes	Yes	Yes	No
N	98,583	12,482	8,517	35,751	5,113	3,103

Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

†Variable included in Triage Test portion of model only

Wait Census Low is omitted category

load variables are removed from the doctor testing equation. Thus, we can safely conclude that across all six columns of Table 18 we see that correcting for potential unobserved correlation only strengthens our conclusion that doctors sometimes reduce testing as crowding increases.

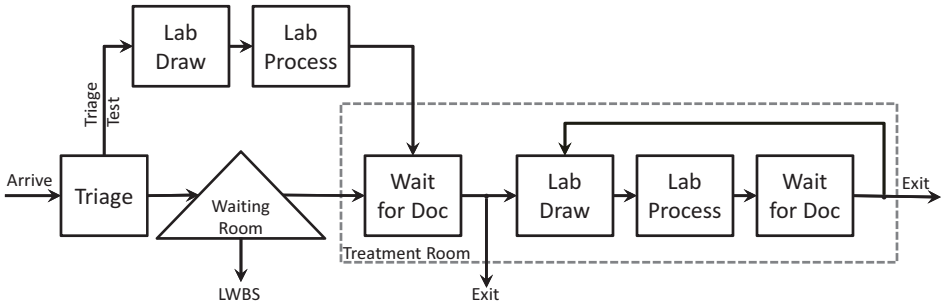
To further check the robustness of our findings regarding the presence of task reduction (Hypothesis 3), we repeat the main study reported in Table 12 with two special subsets of the data. We first test for task reduction for patients that receive no triage tests. Clearly, this is a non-random sample, but it is free of any convoluting effects of doctors responding to triage testing. We find largely the same results as in Table 12 with little evidence of task reduction in the ED while task reduction is present for FT patients in aggregate and for limb pain patients specifically. The second subset we examine is whether abdominal pain and headache patients receive a radiology scan. About 40% of these patients receive a scan, but the scan is ordered by the doctor 99% of the time. Thus, this sample is effectively clear of triage testing treatment bias. We find no evidence of reduced testing under crowding. Taken together, all these robustness checks support or strengthen our main findings regarding Hypothesis 3 that doctors make limited use of task reduction under crowding.

4.8. Simulation

Given our findings of several forms of state-dependent service times in the ED, we are interested in determining what impact these have on performance models. To estimate the impact of the state-dependencies, we build a discrete-event-simulation (DES) model of the ED. Figure 10 diagrams the patient flow in the model. While the model is abstracted from reality, we maintain the essential elements that allow for state-dependent service times, namely the triage testing and doctor testing decisions are state-dependent, and the processing times for Lab Draw and Wait for Doc are state-dependent as well.[7] One additional state-dependency included in the model is the Left Without Being Seen or abandonment

[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.

Figure 10: Patient Flow in Simulation Model



rate. While we do not focus on this phenomenon in this paper, our data clearly exhibits a strong positive correlation between LWBS and waiting room census.

We test three configurations of the model (Table 19). In the first configuration (column 1), all state-dependent variables are included and the model is tuned to match the average performance of our study ED. In the second configuration (column 2), the Speedup and Slowdown state-dependencies are deactivated by fixing all variables at their mean values. In the third configuration (column 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 19: 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 column 2 to column 1 we see that ignoring the Speedup and Slowdown mechanisms leads to a small overestimation of all of the performance measures. Comparing column 3 to column 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 plan-

ning 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 will soon be required to report 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.

4.9. Discussion & Future Work

Prior research has shown that worker-paced service systems tend to exhibit state-dependent service times. In this paper we explore the mechanisms that lead to state-dependent service times whether from a single resource or between multiple resources. We find evidence of both Speedup and Slowdown mechanisms. In our setting, the Slowdown effects tend to dominate in the emergency department, while in the FastTrack, the effects of Slowdown and Speedup balance out.

We find strong evidence of triage-ordered testing being used to reduce in-room service time during periods of crowding in both the ED and the FT. Triage testing saves time by starting tests sooner and allowing at least some of the lab collection and processing time to occur in parallel with the patient waiting time. The main downside to triage testing is the financial cost of unneeded tests. Since neither an insured patient nor the triage nurse directly incur the financial cost, it likely does not weigh heavily on the testing decision. Given the effectiveness of triage testing as a form of Speedup, it is curious that triage testing is not used more regularly, regardless of crowd level. Our findings suggest that hospitals could potentially benefit from increased use of triage testing. Managers should further explore the true costs of over testing at triage and consider incorporating load-based guidelines into triage nurse protocols.

We find evidence of care providers reducing testing orders in the FT when the system is crowded but only limited evidence of this in the ED. In the healthcare setting, task reduction is clearly a double-edged sword. On the 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. (We found no evidence of crowding leading to an increase in 72-hour revisits, a common ED quality metric, in either the ED or the FT.) Determining the "optimal" level of corner cutting 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). As Operations Management researchers, we are satisfied to show that task reduction under load does exist in some circumstances and serves to speed up a service system. Thus, again our work suggests that hospital managers should explore the quality trade-offs of task reduction and should potentially include load-based guidelines in care protocols.

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. While this work focused on server-level state-dependencies, future work should also look at patient-level 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 work load. We hope that incorporation of these mechanisms into future normative models will lead to better understanding and management of similar service systems with high server discretion.

Acknowledgments

This research is partially supported by grants from the Wharton Risk Management and Decisions Processes Center and the Fishman-Davidson Center for Service and Operations Management.

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

The negative binomial logit hurdle 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.

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

CHAPTER 5 : Financial Consequences of Boarding[1]

5.1. Introduction

Emergency Department (ED) crowding has been identified as a public health problem by the Institute of Medicine (Institute of Medicine 2007). When EDs are crowded, patients leave without being seen (LWBS) and some later return for urgent medical needs (Asaro et al. 2007, Rowe et al. 2006, Baker et al. 1991). Ambulance diversion, a hospital's response to crowding, can delay care for time-sensitive diseases, including thombolysis for acute myocardial infarction (Schull et al. 2004). ED "boarding" is one of the major causes of ED crowding, where admitted ED patients spend long periods awaiting inpatient beds (Government Accountability Office 2009, Hoot et al. 2008, Solberg et al. 2003). As boarding increases within an ED, fewer ED resources are available for new patients. This leads to delays in antibiotics for pneumonia and pain control, and higher complication rates (Pines et al. 2007, Pines and Hollander 2008, Fee et al. 2007, Pines et al. 2009). One study estimated that 15% of the overall time spent in U.S. EDs by patients boarding (Carr et al. 2010). Boarding itself is associated with higher medical error rates, and has proven hazardous for patients admitted to intensive care settings (Carr et al. 2007a, Chalfin et al. 2007, Liu et al. 2009, Kulstad et al. 2010).

A recent discussion has begun in academic medical journals and the lay press about whether the practice of ED boarding may actually increase a hospital's revenue (Meisel and Pines 2008, Goldstein 2008). Overflow capacity in ED hallways can be used as a temporary holding area, allowing the hospital to operate at higher occupancy than it has in licensed beds. Concerns have been raised that hospitals have perpetuated ED boarding because of insufficient economic incentive to eliminate it. However, data have been mixed. Some studies suggest that the economic impact of ED crowding and diversion is lost revenue as

[1] This chapter is reprinted from Pines, Jesse M., Robert J. Batt, Joshua A. Hilton, and Christian Terwiesch. "The financial consequences of lost demand and reducing boarding in hospital emergency departments." Annals of Emergency Medicine 58, no. 4 (2011): 331-340. with permission from Elsevier

patients LWBS and ambulance patients are directed elsewhere (Lucas et al. 2009, Falvo et al. 2007). Others conclude that ED crowding and diversion maximizes revenue because ED admissions generate less revenue than non-ED admissions (McHugh et al. 2008). In a situation where there is plentiful demand for both ED and non-ED admissions, crowded EDs may allow hospitals to prioritize inpatient beds for elective (non-ED) patients from whom hospitals can collect higher reimbursement (Pines and Heckman 2009, Handel et al. 2010). During weeks of high diversion, one hospital collected $265,000 more in revenue than during weeks of low diversion (Handel and John McConnell 2009). The key tradeoff lies in balancing increased revenue from capturing lost ED demand (lowering LWBS and diversion) versus the potential lost revenue from reducing non-ED admissions to open capacity to serve higher ED demand.

We examined the tradeoff between the higher revenue from capturing ED demand versus potential losses from reducing non-ED admissions by simulating what may happen to hospital revenues if average boarding is reduced by an hour. We also determined how different bed management policies for reducing non-ED admissions to accommodate additional ED demands would impact hospital revenue. Our overall goals were to determine if reducing boarding increases or decreases hospital revenue, and how a hospital could potentially better manage non-ED demand to ensure that reducing boarding would result in increased revenue.

5.2. Methods

5.2.1. Study Design, Setting, and Selection of Participants

A stepped approach was used to estimate the revenue implications of the balance between reducing boarding and the need to reduce non-ED admissions to accommodate new ED demand (LWBS and diversion). We first calculated the expected value of lost ED demand, specifically the expected revenue from serving additional LWBS patients and patients who were diverted to other hospitals. We then calculated the expected value of revenue change from reducing the mean boarding time by one-hour using two methods: (1) a financial

model informed by the results from regression analyses and (2) a discrete-event simulation model to validate and extend the first analysis by simulating how specific types of inpatient bed management policies (with regard to reducing the inflow of non-ED admissions) may increase revenue (or not).

In the simulation, we calculated the percent reduction in non-ED admissions necessary to serve the increased number of ED admissions that would result from reducing boarding using two potential bed management policies. First, we estimated how reducing non-ED admissions by a fixed proportion would impact overall revenue. This was termed a "static" policy. Next, we estimated how various types of "dynamic" management policies impacted revenue. Dynamic policies were defined by two parameters – the proportion reduction in non-ED admissions and the specific trigger point (i.e. the bed number at which a reduction would be deployed). Various static and dynamic bed management strategies were tested to determine which allowed the ED to maintain current service levels and which, if any, resulted in higher overall revenue at the hospital level.

The data included for the calculations were all ED patients registered and all non-ED patients (direct admissions and transfers) admitted to a single, inner-city teaching hospital over a two-year period (FY 2007-8). Excluded were patients admitted to inpatient rehabilitation, psychiatry, and labor and delivery because they are not seen in the study ED and do not compete directly with ED patients for inpatient beds. Also included were actual data on ambulance diversion (separated by medical and trauma) over the study period. LWBS patients were included if they were triaged, and each had a triage level which was used for analysis. Patients who left before treatment complete or left against medical advice were included as treated and discharged patients because they still resulted in revenue.

For each ED patient, we used data on arrival date, time, and mode (ambulance v. non-ambulance), triage level, disposition, and actual revenue received. We used timestamps for patient movement through the ED: earliest arrival, placement in treatment room, inpatient bed request, and departure from the ED. From these timestamps, durations were calculated

for wait time (earliest arrival until room placement), service time (room placement to departure for outpatients or bed request for admitted patients), and boarding time (bed request to departure for admitted patients). Timestamps were obtained directly from the electronic medical record, which stores timestamps in real-time as a regular part of ED workflow.

5.2.2. Outcome Measures

The main outcome was direct revenue. Indirect revenue, including federal payments to support resident education, was not included. We did not include direct or indirect costs, and assumed hospital costs were largely fixed (Roberts et al. 1999). Therefore, we did not calculate actual contribution margins or profitability, because cost allocation methods vary widely between hospitals. Therefore, changes in revenue served as a proxy for changes in hospital profitability. Revenue was classified by the patient type, not by where the revenue charge was incurred. For example, the total revenue generated from a patient visit that started in the ED and was then admitted for three days would be classified as ED-admission revenue.

5.2.3. Primary Data Analysis

To quantify the revenue lost by LWBS and diversion, we estimated the expected value of LWBS patients and both medical and trauma diversion hours. The expected dollar value of a LWBS patient was estimated by the following weighted sum:

$$E[LWBS] = \sum_{i=1}^{4} \Pr\left(TriageLevel_i\right) \left[\Pr\left(admit_i\right) E\left[\text{Revenue}_{\text{admit}_i}\right] + \left(1 - \Pr\left(admit_i\right)\right) E\left[\text{Revenue}_{\text{out}_i}\right]\right] \tag{5.1}$$

Triage level probabilities were calculated from observed LWBS patients. Since there were no data on admission rates for LWBS patients had they remained for treatment, we used the admission rates for ambulatory patients, conditional on triage level, as a proxy for LWBS admission rates since the vast majority of LWBS patients are ambulatory. However, it is

possible that due to self-selection, LWBS patients would be less likely to be admitted than those who stayed. We therefore conducted a sensitivity analysis on the LWBS admission rate and reported results for LWBS admission rates that were assumed in the financial model to be half that of the observed population. However, later in the simulation, we tested an admission rate of zero for LWBS and the triage-level adjusted rate because of the limitations of the simulation software. The annual lost revenue from LWBS was obtained by multiplying the expected value of a single LWBS patient by the number of LWBS patients.

The value of an hour of medical diversion was calculated as the product of the expected revenue of a single medical ambulance arrival and the expected number of medical arrivals per hour. The value of a medical arrival was calculated similar to LWBS patients except that admission probabilities and expected revenues were estimated from medical arrivals. The expected ambulance arrival rate was estimated by dividing the number of medical arrivals by the number of hours the hospital was not on medical diversion during the study period. Annual lost revenue from medical diversion was calculated as the product of the expected value of an hour of medical diversion and the number of hours the hospital was on medical diversion in a given year. The value of trauma diversion was estimated similarly from trauma ambulance arrivals.

Next, we estimated the effect of boarding on revenue through two methods: first with a financial model informed by regression and second with discrete-event simulation. The regression method used ordinary least squares regression and drew on the relationship between the mean daily boarding and the number of daily LWBS patients. Since the number of ED arrivals (i.e. daily demand) influences both boarding times and LWBS, we used the following model:

$$CountLWBS\_Day_t = \beta_0 + \beta_1 AvgBoarding + \beta_2 CountArrivals_t + \varepsilon_t \qquad (5.2)$$

We used similar models for hours of medical and trauma diversion, replacing the count of

LWBS with the number of hours of diversion per day.

Because of the relatively low explanatory power in the relationship between boarding and LWBS and diversion (R^2 of 0.43, 0.25, and 0.24 respectively for LWBS, medical diversion, and trauma diversion), discrete-event simulation was used to validate the estimates of the changes in boarding on revenue. Simulation was also used to extend the analysis to estimate how the increased inpatient load from the new ED demand would impact overall hospital operations; specifically, the potential reduction in non-ED admissions necessary to serve the new inpatient load generated by more ED admissions. With the simulation model, we created a virtual ED and hospital by using patient-level data to estimate probability distributions of patient flow. The model permitted us to change a parameter (i.e. mean boarding time), and observe the effects on revenue.

The discrete-event simulation model had three ED arrival streams: medical ambulance, trauma ambulance, and ambulatory (Figure 11). Each stream was an independent Poisson arrival process estimated from data and designed to mirror ED operations. To simulate LWBS behavior, we drew on abandonment and impatience models from queuing theory (Gans et al. 2003). Each patient was assigned a maximum waiting time drawn from a probability distribution. A Weibull distribution with shape parameter greater than one was used to simulate increasing impatience (Gross et al. 2008).

Diversion was triggered by queue length. After crossing a trigger point, the relevant arrival stream was diverted from the ED for four hours (which mirrored study hospital policy). After time expired, the arrival stream reopened if the queue length was below the trigger point, otherwise another four hours of diversion occurred.

Most parameters were estimated directly from the study data (Table 20). However, several parameters could not be directly estimated: abandonment time distributions, diversion triggers, and number of beds. Therefore, we used sensitivity analysis and an evolutionary optimizer to tune the model to match the real results, and independence was verified between

Figure 11: Discrete-event model of the ED

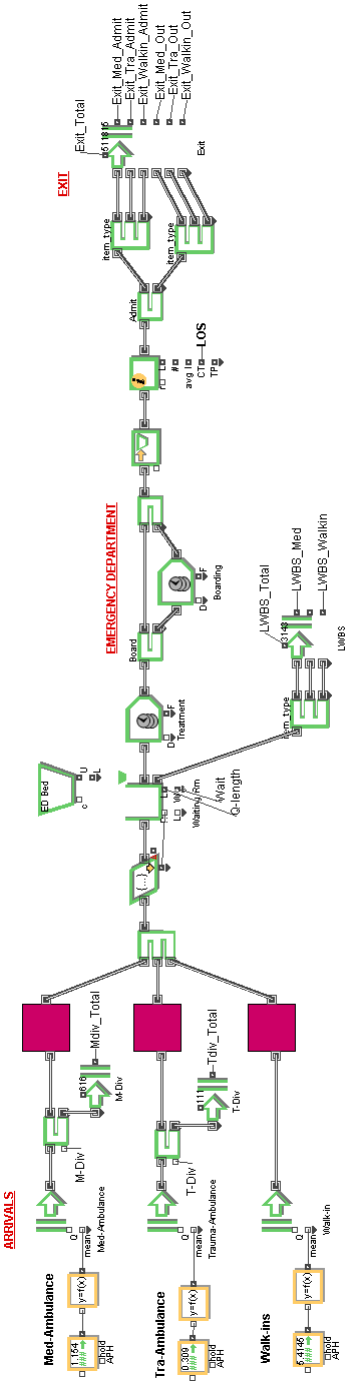


Table 20: Descriptive statistics of the study population (fiscal year 2007 to 2008, by arrival type)

Arrival Type	**Medical Ambulance**	**Trauma Ambulance**	**Walk-in**
Count	17,856	4,914	92,462
Median Age, yr.	46	38	36
% Male	46%	70%	39%
% White	19%	32%	21%
% Black	70%	55%	66%
Arrival Rate, patients/hr. (SE)	1.12 (0.02)	0.30 (0.01)	5.41 (0.03)
Admit Probability	34%	57%	18%
Boarding Probability	96%	6%	94%
Service Time for Admitted Patients			
Distribution Type	Gamma	Gamma	Gamma
Scale	2.10	2.60	1.99
Shape	1.73	1.60	2.02
Mean (SD)	3.63 (2.74)	4.16 (4.33)	4.02 (3.07)
Median (IQR)	3.03 (3.06)	3.27 (3.11)	3.33 (3.25)
Service Time for Outpatients			
Distribution Type	Gamma	Gamma	Gamma
Scale	3.19	5.96	3.53
Shape	1.59	1.24	1.06
Mean (SD)	5.07 (5.17)	7.39 (6.74)	3.74 (5.00)
Median (IQR)	3.85 (3.75)	4.98 (6.55)	2.57 (3.33)
Boarding Time			
Distribution Type	Weibull	Weibull	Weibull
Scale	3.64	3.19	3.52
Shape	0.955	0.936	0.891
Mean (SD)	3.73 (4.71)	3.30 (3.77)	3.75 (5.18)
Median (IQR)	2.34 (2.87)	2.03 (3.11)	2.19 (2.88)
Mean Time in ED, hr. (SD)	6.2	6.5	6.0
% LWBS	2%	0%	8%
% of time on Diversion	9%	7%	N/A
Mean Revenue, $ (SD)	4,672 (12,350)	16,529 (36,370)	2,530 (9,849)
Median Revenue, $ (IQR)	497 (6,031)	5,412 (15,351)	334 (742)

SE, Standard Error; SD, Standard Deviation; IQR, Interquartile Range; LWBS, left without being seen.

simulation samples by checking for autocorrelation with the Portmanteau test which found no significant autocorrelation. The simulation compared a base model with a model where the mean boarding time was reduced by one-hour (by reducing the scale parameter in a Weibull distribution). When comparing simulations, we used a paired-t confidence interval (Law 2007). To estimate the revenue effects from the changed model, the estimated change in the number of patients served per day by type was multiplied by the expected revenue for each given patient type.

Next, we estimated the reduction in non-ED admissions that would be required if boarding was reduced. Reducing boarding creates additional demand for inpatient beds in two ways: (1) ED-admitted patients move to inpatient beds earlier, (2) lower boarding reduces lost demand (diversion and LWBS), increasing ED admissions. The required reduction of non-ED admissions depends upon the degree to which the overall hospital is capacity-constrained (i.e. the number of beds available on any given day). Consider three scenarios:

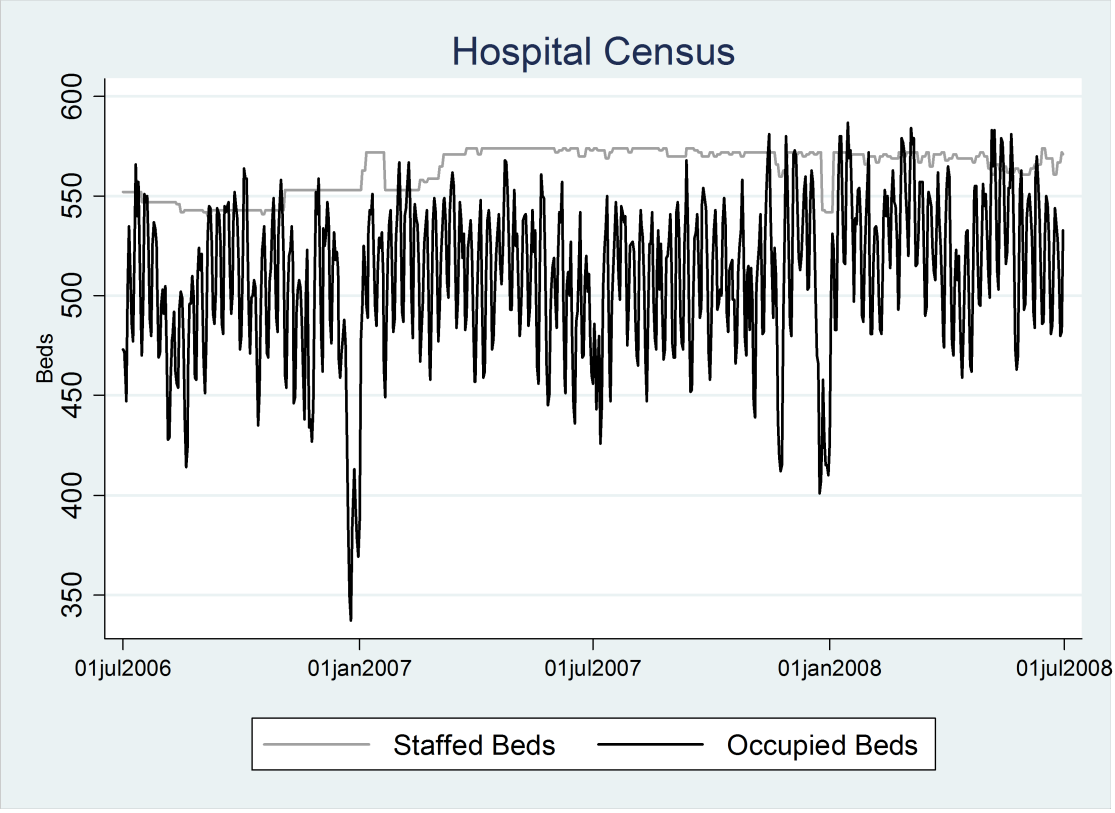
1. Inpatient beds are not capacity-constrained. In this scenario, the hospital can serve the new demand without any cancellations or reductions. It is implicit that boarding is not directly caused by a lack of inpatient beds (Hoot et al. 2008), but results from other inefficiencies.

2. Inpatient beds are completely capacity-constrained. In this case, each patient-hour of increased ED demand from lower boarding and higher ED admissions would require elimination of a patient-hour of non-ED admission.

3. Inpatient beds are periodically capacity-constrained. If the hospital is not always at capacity, only a portion of new demand would necessitate reductions of non-ED admissions.

The first two scenarios serve as boundaries to the potential financial outcomes from reducing boarding (i.e. best and worst case scenario). The third scenario is of primary interest, and simulation was used to test how various non-ED admission reduction policies would allow

Figure 12: Hospital census during the study period



the hospital to serve the increased ED demand and maximize estimated revenue. Policies tested included: 1) a simple across-the-board reduction of the non-ED admission rate (i.e. a static model) and 2) dynamic policies that actively scaled back non-ED admissions by specific proportions only when the hospital was above a given census trigger point.

To test these policies, we assumed a hospital capacity of 565 beds which was the average staffed-bed capacity of the study hospital (Figure 12). Capacity data was calculated from actual arrivals and departures of ED and non-ED admissions. Staffed-beds were determined from random daily snapshots of hospitals' staffed-beds using Navicare software (Hill-Rom, Batesville, IN), the management tracking system for staffed-beds and census.

We first determined in a base-case model the proportion of ED admissions who boarded in the ED directly due to capacity constraints (i.e. no appropriate bed was available). This served

as the service level target for all potential scheduling policies. Mean boarding time was then reduced by one-hour and the service level for a total of 80 potential policies was measured (10 static and 70 dynamic policies). The first question was whether a policy matched or exceeded the service level, determined as a policy where no additional inpatient capacity would be needed. We then simulated the daily increase or decrease in daily revenue from the increase in ED demand and reduction policy for non-ED admissions. The objective was to find the non-ED admission policy or set of policies that would both match or exceed the target service levels and maximize net revenue gains under the reduced boarding scenario.

Analyses were performed using Microsoft Excel 2007 (Microsoft Corporation, Redmond, WA), Stata 10 (Stata Corp, College Station, TX), ExtendSim 8 (Imagine That Inc., San Jose, CA), and JMP 8.0 (SAS Institute Inc., Cary, NC). This study received approval from the institutional review board.

5.3. Results

A total of 92,456 ED outpatients, 25,753 ED admissions, and 36,393 non-ED admissions were used for analysis over two-years. Median hospital length of stay for ED and non-ED admissions was 3 days. Mean revenue for ED outpatients was $647, ED admissions $2,268 per patient-day, and non-ED admissions $4,118 per patient-day (Table 21).

There were 3,186 LWBS encounters during FY2007 and 3,845 during FY 2008. The expected value for one LWBS patient was $1,096, assuming admission of LWBS patients occurs at half the rate of the observed ambulatory population, conditional on triage level. In sensitivity analysis, when all LWBS patients were outpatients, the expected value was $478 and when LWBS patients were admitted at the same rate as those that stayed by triage level, the expected value was $1,714. Treating all LWBS patients (assuming an admission rate of $\frac{1}{2}$ the observed ambulatory rate) would have resulted in an additional $3.5 million in revenue in FY2007 and $4.2 million in revenue in FY2008.

There were 618 and 1,020 medical diversion hours and 479 and 794 trauma diversion hours in

Table 21: Descriptive statistics of the study population in a single hospital during a 2-year period (fiscal year 2007 to 2008)

Variables	ED Outpatients	ED Admissions	Non-ED Admissions
Patient Count (%)	92,456 (60%)	25,753 (17%)	36,393 (24%)
Revenue, $, millions (%)	59.8 (5%)	338.7 (26%)	929.2 (70%)
Mean Length of Stay, days (SD)	N/A	5.8 (9.1)	6.2 (9.4)
Median Length of Stay, days (IQR)	N/A	3 (4)	3 (5)
Mean Revenue/Patient per Day, $	647	2,268	4,118
Median Revenue/Patient per Day, $ (IQR)	226 (425)	2,242 (1,966)	3,556 (5,482)

N/A, Not Applicable.

FY2007 and FY2008 respectively. During off-diversion times, there were 1.2 non-ambulance arrivals per hour for medical patients and 0.3 ambulance arrivals per hour for trauma patients. The expected revenue for a medical ambulance arrival was $4,670 and the expected revenue for a trauma arrival was $16,526. The expected lost revenue from one hour of medical diversion was $5,388 and the expected lost revenue from each hour of trauma diversion was $5,110. Medical diversion resulted in forgone revenue of $3.3 million and $5.5 million in FY2007 and FY2008. Trauma diversion resulted in $2.4 million and $4.1 million in forgone revenue in FY2007 and FY2008. The overall estimated lost revenue from lost demand was $9.3 million for FY2007 and $13.8 million for FY2008.

For the 25,753 ED admissions in FY2007 and FY2008, the mean boarding time was 3.7 hours (standard deviation [SD] 5.2 hours), and median boarding time was 2.2 hours (Interquartile range [IQR] 1.1 – 4.1). A one hour change in average boarding time was associated with a change of 1.1 (95% CI 0.9 - 1.3) patients per day who LWBS. Regression analyses found that a one-hour reduction in average boarding time was associated with a 1.2 hours per day (95% CI 0.9-1.5) reduction in medical diversion hours and 0.7 hours per day (95% CI 0.5-1.0) in trauma diversion hours. Using the estimated values of LWBS and diversion, a

Table 22: Changes in the number of patients served with 1-hour reduction in mean boarding and expected revenue

Variables	Change in Mean Patients Served (SE)	Expected Revenue per Patient, $ (SE)
ED Medical Ambulance Admission	0.35 (0.02)	12,296 (235)
ED Trauma Ambulance Admission	0.11 (0.01)	24,352 (856)
ED Ambulatory Admission	0.00 (0.01)	11,704 (159)
ED Medical Ambulance Outpatient	0.85 (0.03)	723 (32)
ED Trauma Ambulance Outpatient	0.09 (0.01)	6,361 (319)
ED Ambulatory Outpatient	2.81 (0.04)	499 (6)

one-hour reduction in average boarding time would increase revenue by $11,301 per day. This estimate ranged from $10,628 to $11,974 as the LWBS admission rate assumption was varied from 0% to the observed ambulatory admission rate. In the simulation, if all LWBS patients were outpatients, this would result in $9,693 increased revenue per day, or $3.5 million per year. When LWBS admission rates mirror ambulatory admission rates, reducing boarding by an hour would increase revenue by $13,298 per day or $4.9 million per year. The estimated values used in the simulation for each patient type based on the study data are listed in Table 22.

A one-hour reduction in mean boarding led to an increase in inpatient bed demand of 4.4 bed-days per day (1.3 days for reducing boarding and 3.1 days for accommodating additional ED admissions). Assuming that inpatient beds are never capacity-constrained, reducing boarding by an hour would increase hospital revenue by $3.5 million per year and require no reduction in non-ED admissions. Assuming that inpatient beds are always capacity-constrained, the new inpatient demand would necessitate non-ED admission cancellations worth $18,172 per day. The hospital would therefore experience a net revenue reduction of $8,479 per day or $3.1 million per year if it reduced mean boarding time by one hour in a completely capacity-constrained situation.

In the scenario that inpatient beds are intermittently capacity constrained, the financial

Table 23: Non-ED admission policy comparison for net change in revenue caused by 1-Hour average ED boarding reduction, in which LWBS patients are all ED outpatients

Non-ED Admission Reduction Percentage	Static Policy: No Trigger	Dynamic Policies – Trigger Census at which reduction is implemented						
		530	535	540	545	550	555	560
1%	$ (2,993)	$ 5,112	$ 6,060	$ 6,939	$ 7,717	$ 8,302	$ 8,768	$ 9,140
2%	$ (15,679)	$ 1,156	$ 2,889	$ 4,614	$ 5,974	$ 7,087	$ 7,971	$ 8,637
3%	$ (28,365)	$ (2,469)	$ 120	$ 2,339	$ 4,452	$ 5,933	$ 7,198	$ 8,190
4%	$ (41,051)	$ (5,984)	$ (2,466)	$ 719	$ 2,869	$ 5,029	$ 6,533	$ 7,719
5%	$ (53,738)	$ (9,121)	$ (4,842)	$ (1,049)	$ 1,713	$ 4,104	$ 5,912	$ 7,418
6%	$ (66,424)	$ (12,094)	$ (6,961)	$ (2,885)	$ 917	$ 3,315	$ 5,415	$ 7,078
7%	$ (79,110)	$ (14,174)	$ (9,179)	$ (4,735)	$ (479)	$ 2,512	$ 4,768	$ 6,649
8%	$ (91,796)	$ (16,698)	$ (11,433)	$ (5,983)	$ (1,315)	$ 1,742	$ 4,121	$ 6,293
9%	$ (104,483)	$ (19,100)	$ (12,700)	$ (7,505)	$ (2,670)	$ 1,072	$ 4,029	$ 6,069
10%	$ (117,169)	$ (21,099)	$ (14,474)	$ (9,146)	$ (3,803)	$ 510	$ 3,412	$ 5,694

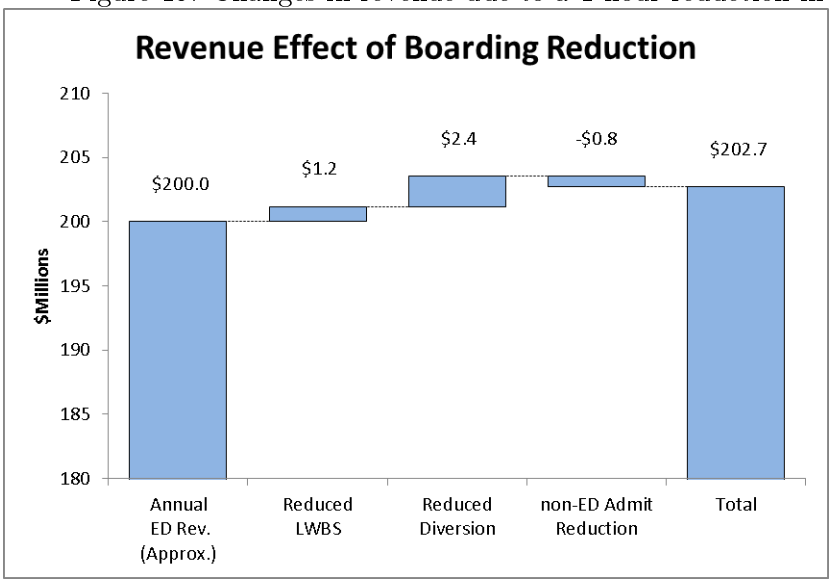
Max hospital capacity is 565 beds. Compares static policies in which non-ED admissions are reduced across the board versus dynamic policies in which non-ED admissions are reduced by specific percentages at specific trigger censuses.
Values represent increase (decrease) in daily revenue from the policy.
Shaded cells represent policies that require additional bed capacity.
All other values can be achieved with no increase in bed capacity.

Table 24: Non-ED admission policy comparison for net change in revenue caused by 1-Hour average ED boarding reduction, in which LWBS patients are admitted at rates mirroring those of patients who stayed for care.

Non-ED Admission Reduction Percentage	Static Policy: No Trigger	Dynamic Policies – Trigger Census at which reduction is implemented						
		530	535	540	545	550	555	560
1%	$ 612	$ 9,540	$ 10,433	$ 11,242	$ 11,793	$ 12,357	$ 12,695	$ 12,963
2%	$ (12,074)	$ 6,140	$ 7,758	$ 9,327	$ 10,432	$ 11,454	$ 12,181	$ 12,694
3%	$ (24,760)	$ 3,244	$ 5,631	$ 7,518	$ 9,429	$ 10,742	$ 11,785	$ 12,436
4%	$ (37,447)	$ 403	$ 3,592	$ 5,895	$ 8,344	$ 9,955	$ 11,367	$ 12,122
5%	$ (50,133)	$ (2,039)	$ 1,629	$ 4,503	$ 7,444	$ 9,654	$ 10,782	$ 11,918
6%	$ (62,819)	$ (4,085)	$ 136	$ 3,379	$ 6,186	$ 8,777	$ 10,366	$ 11,842
7%	$ (75,505)	$ (6,210)	$ (1,656)	$ 2,066	$ 5,644	$ 8,229	$ 10,076	$ 11,485
8%	$ (88,191)	$ (8,146)	$ (3,398)	$ 879	$ 4,617	$ 7,891	$ 10,009	$ 11,324
9%	$ (100,878)	$ (9,793)	$ (4,512)	$ 211	$ 3,663	$ 7,421	$ 9,690	$ 11,145
10%	$ (113,564)	$ (12,185)	$ (6,040)	$ (1,233)	$ 2,935	$ 6,913	$ 9,252	$ 10,911

Max hospital capacity is 565 beds. Compares static policies in which non-ED admissions are reduced across the board versus dynamic policies in which non-ED admissions are reduced by specific percentages at specific trigger censuses.
Values represent increase (decrease) in daily revenue from the policy.
Shaded cells represent policies that require additional bed capacity.
All other values can be achieved with no increase in bed capacity.

Figure 13: Changes in revenue due to a 1 hour reduction in mean boarding time



results depend on the non-ED admission policy. Tables 23 and 24 demonstrate the daily change in revenue under different static and dynamic bed management policies. In the case where LWBS patients are considered outpatients, 1% reduction in ED admissions or lower did not meet current service levels and there were no static policies that resulted in increased revenue for the hospital. The 70 dynamic policies tested ranged from trigger censuses of 530 to 560 beds and a 1% to 10% reduction. Of those, 55 met current service levels and 35 policies would result in higher revenue. The optimal strategy was a 5% reduction in non-ED admissions at 560 beds resulting in $7,418 higher revenue per day or $2.7 million per year (Figure 13). In the case where LWBS patients were admitted at the ambulatory admission rate, all static policies met current service levels and a 1% reduction in non-ED admissions was the only higher revenue policy resulting in $612 in greater revenue per day or $223 thousand per year. Of the 70 dynamic policies tested, 25 met current service levels, and of those, 14 would result in higher revenue. The optimal strategy was an 8% reduction at 555 beds, resulting in $10,009 or $3.6 million per year.

5.4. Limitations

A major limitation of our study is that we used data from a single hospital. Other hospitals with different processes may experience different revenue effects than we found (Henneman et al. 2009). For example, Massachusetts hospitals that are by law no longer permitted to go on diversion may experience smaller gains from reducing boarding than hospitals that regularly divert ambulances. We found also that trauma arrivals resulted in considerably higher revenue than medical arrivals, which may not be seen in other hospitals. This may be explained by local factors, such as negotiated agreements with payers, or the fact that in Pennsylvania (the site of the study hospital), a law requires 100% reimbursement of charges for worker's compensation trauma victims. In addition, because in this inner city hospital, ED patients were more likely to be uninsured and have Medicaid insurance than non-ED patients, there was more than 1.5 fold difference between ED admission and non-ED admission revenue. In hospitals with more balanced payer-mixes between ED and non-ED admissions, we would expect the potential revenue gains from reducing boarding to be higher. In addition, because this hospital was an inner-city hospital, the potential revenue losses from diversion would be expected to be higher because of the higher likelihood of penetrating trauma victims requiring operative management.

Our model was also simplistic in that we assumed bed pooling between specific types of beds (i.e. pooling floor, telemetry and intensive care), which may not reflect policies in other hospitals imposing stricter rules about segregating service lines within units. Restrictions on bed pooling would serve to reduce the gains from lower levels of boarding. In addition, we assumed the staffed-bed capacity to be in our model fixed which was not completely reflective of reality (Figure 12). Staffed-bed variability may be even greater in many hospitals, which may result in less unfilled staffed occupancy. We also made an assumption that hospital expenses are largely fixed and we used revenue as our main outcome. The degree to which hospital staffing would need to be increased to accommodate the increased demand, particularly if more expensive temporary staffing was used, may lower our estimates of the

financial benefits of reducing boarding. Lastly, we did not directly calculate how reducing ED crowding and boarding may impact outcomes. Given studies that have demonstrated higher medical error rates and complications associated with crowding (e.g., Pines et al. 2007, Pines and Hollander 2008, Fee et al. 2007), it is likely that the impact on outcomes such as lower complications and shorter lengths of stay would serve to further increase hospital revenues if boarding is reduced. It is also possible that reducing boarding may have downstream effects, such as changing the likelihood of an emergency physicians' decision to admit.

5.5. Discussion

Studies on the revenue impacts of boarding have shown mixed results (e.g., Lucas et al. 2009, Falvo et al. 2007, McHugh et al. 2008). The potential gains from reducing boarding have been estimated in some studies, while in others direct comparisons between ED admissions and non-ED admissions have been made that have shown that ED admissions are less profitable than non-ED admissions in broad populations. No studies have directly assessed the tradeoff between potentially lost revenue from LWBS and diversion and the degree to which any reduction of boarding would necessitate lower numbers of financially attractive, non-ED admissions. We advance the understanding of this balance by demonstrating the potential revenue gains or losses under various conditions from reducing boarding by one hour using data from a single hospital. Specifically, we demonstrate how overall hospital revenue can change dramatically based on the different policies employed to manage hospital capacity by selectively reducing non-ED admissions on higher demand days to allow for lower ED boarding times.

The two types of policies tested were static - reducing the average number of non-ED admissions per day – and dynamic – using active scheduling to strategically reduce non-ED admissions on higher demand days. In the case where LWBS patients were outpatients, there was no static policy that allowed the ED to reduce boarding, maintain current service levels, and generate revenue gains, while in the case where LWBS patients are admitted at

the ambulatory rate, a 1% across-the-board reduction was marginally revenue positive. This indicates that across-the-board reductions in non-ED admissions to improve the functioning of the ED are likely not a financially attractive strategy for hospital managers.

However, many dynamic policies allowed for a maintenance of the same non-ED admission rate, as long as the hospital census was below a given trigger point. Once the trigger point was reached, non-ED admissions would be reduced by a given percentage until the census dropped below the trigger point. Assuming that LWBS patients are outpatients, the optimal dynamic policy called for a 5% reduction in non-ED admissions when the census reached 560, while assuming their admission rate is the same as their triaged counterparts who stayed for care, the optimal policy would be an 8% reduction when the hospital census reached 555. During the study period, the hospital admitted about 50 non-ED patients a day, so a 5-8% reduction would require cancellation of approximately 2-4 non-ED patients when the trigger census is reached. This assumes that patients are cancelled and their revenue is lost forever, therefore if patients could be rescheduled rather than lost, the revenue estimates may underestimate the net revenue change.

Our results also show that a wide range of dynamic policies are acceptable and achieve relatively similar results. Hospital managers may have various reasons to select a particular policy (i.e. one that favors a lower trigger or a lower reduction rate). There is also a tradeoff that certain trigger rates would require hospitals to spend more days in a "non-ED admission reduction mode." Higher administrative costs, customer service concerns, or the response from inpatient services who gain more revenue from non-ED admissions may also play into which particular active management plan is chosen.

This study also provides evidence that calls into question the commonly held belief that boarding is largely caused by a lack of inpatient beds (Henneman et al. 2009). In the simulation, increases in ED admissions were accommodated on most days without any change to non-ED admissions and the staffed-beds were mostly higher than the hospital census (Figure 12). In fact, reducing boarding rarely pushes existing patients out, assuming that

the hospital is making best use of its staffed-space, which may not be the case. Under the various policies tested, reducing non-ED admissions was required only 3% - 20% of the time, suggesting that much of observed ED boarding times may not have been cause by a lack of physical beds, but rather by other inefficiencies in the system that slow transitions of care between hospital units, or requirements that specific units house specific types of patients (i.e. the gastroenterology patients can only be on one hospital unit) with little pooling between similar types of beds. Future studies should in managing hospital capacity should study the impact of pooling, and other strategies to better balance non-ED admissions to reduce artificial flow variability through load-leveling (i.e. surgical schedule smoothing).

Several aspects of this calculation make this study generalizable and not generalizable to other U.S. hospitals. The findings would be most generalizable to other large, high-volume, teaching hospitals because they would be likely to experience similar variability in occupancy, demonstrated by large swings in census that frequently go below peak capacity. This would be true particularly in those that have not employed load-leveling of non-ED admission schedule, as was the case in the study hospital. However, in hospitals without the same levels of boarding, LWBS, and diversion, our results may be less applicable. This may be the case in hospitals with no diversion policies or those that make better use of staffed beds.

In summary, we found that ED boarding leads to unfilled patient need – as measured by ambulance diversion and walk-away rates – and large potential losses in hospital revenue. We also demonstrate that the potential revenue impacts of reducing boarding is highly dependent on how a hospital manages the variability in bed capacity in a single inner-city, teaching hospital. Specifically, how the hospital chooses to handle inpatient bed management strategies is vital. How non-ED admissions are reduced to accommodate new demand is the primary driver of whether reducing boarding increases hospital revenues or not. We identified several dynamic admissions policies for non-ED patients that could serve higher demand for ED admissions with minimal effect on non-ED patients and lead to a net revenue gain of \$2.7– 3.6 million per year.

Acknowledgments

The authors acknowledge the many participants who helped us carry out this study, including Christian Boedec, John Heckman, MBA, Joshua A. Isserman, MS, Scott Lorch, MD, and Evan Fieldston, MD.

: Bibliography

ACEP. 2012. Publishing wait times for emergency department care American College of Emergency Physicians. http://www.acep.org/clinical—practice-management/publishing-wait-times-for-emergency-department-care,-june-2012.

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.

Aksin, Zeynep, Baris Ata, Seyed Emadi, Che-Lin Su. 2012. Structural estimation of callers' delay sensitivity in call centers. *Working Paper* .

Alizamir, Saed, Francis deVericourt, Peng Sun. 2011. Diagnostic accuracy under congestion. *Working Paper* .

Allon, Gad, Achal Bassamboo, Itai Gurvich. 2011. We will be right with you: Managing customer expectations with vague promises and cheap talk. *Operations Research* **59**(6) 1382–1394.

Andersen, Arne S, Petter Laake. 1987. A model for physician utilization within 2 weeks: analysis of norwegian data. *Medical care* **25**(4) 300–310.

Argon, Nilay Tanık, Serhan Ziya. 2009. Priority assignment under imperfect information on customer type identities. *Manufacturing & Service Operations Management* **11**(4) 674–693.

Armony, Mor, Shlomo Israelit, Avishai Mandelbarum, Yarvin N Marmor, Yulia Tseytlin, Galit B. Yom-Tov. 2012. Patient flow in hospitals: A data-based queuing-science perspective. *Working Paper* .

Armony, Mor, Nahum Shimkin, Ward Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57**(1) 66–81.

Armony, Mor, Amy R. Ward. 2010. Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research* **58**(3) 624–637.

Aronsky, Dominik, Ian Jones, Kevin Lanaghan, Corey M Slovis. 2008. Supporting patient care in the emergency department with a computerized whiteboard system. *Journal of the American Medical Informatics Association* **15**(2) 184–194.

Asaro, Phillip V, Lawrence M Lewis, Stuart B Boxerman. 2007. Emergency department overcrowding: analysis of the factors of renege rate. *Academic Emergency Medicine* **14**(2) 157–162.

Asplin, B, FC Blum, RI Broida, et al. 2008. Acep task force report on boarding. *Emergency Department Crowding: High-Impact Solutions* .

Asplin, Brent R, David J Magid, Karin V Rhodes, Leif I Solberg, Nicole Lurie, Carlos A Camargo Jr. 2003. A conceptual model of emergency department crowding. *Annals of emergency medicine* **42**(2) 173–180.

Ata, Barış, Jan A Van Mieghem. 2009. The value of partial resource pooling: Should a service network be integrated or product-focused? *Management Science* **55**(1) 115–131.

Baccelli, F., G. Hebuterne. 1981. On queues with impatient customers. *Performance* 159–179.

Baker, David W, Carl D Stevens, Robert H Brook. 1991. Patients who leave a public hospital emergency department without being seen by a physician. *JAMA: the journal of the American Medical Association* **266**(8) 1085–1090.

Batt, Robert J., Sergei Savin, Christian Terwiesch. 2013. Throughput in the emergency department. *Working Paper* .

Batt, Robert J., Christian Terwiesch. 2013. Doctors under load: An empirical study of state-dependent service times in emergency care. *Working Paper* .

Baumann, Michael R., Tania D. Strout. 2005. Evaluation of the emergency severity index (version 3) triage algorithm in pediatric patients. *Academic Emergency Medicine* **12**(3) 219–224.

Berry Jaeker, Jillian, Anita L. Tucker. 2012. Hurry up and wait: Differential impacts of congestion, bottleneck pressure, and predictability on patient length of stay. *Working Paper* .

Bertoty, David A, Michele L Kuszajewski, Eric E Marsh. 2007. Direct-to-room: one department's approach to improving ed throughput. *Journal of Emergency Nursing* **33**(1) 26–30.

Bitran, Gabriel R., Juan-Carlos Ferrer, Paulo Rocha e Oliveira. 2008. Managing customer experiences: Perspectives on the temporal aspects of service encounters. *Manufacturing & Service Operations Management* **10**(1) 61–83.

Boger, Elisa. 2003. Electronic tracking board reduces ed patient length of stay at indiana hospital. *Journal of Emergency Nursing* **29**(1) 39–43.

Brandt, Andreas, Manfred Brandt. 2002. Asymptotic results and a markovian approximation for the m(n)/m(n)/s+gi system. *Queueing Systems* **41** 73–94.

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

Brown, Lawrence, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, Linda Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* **100**(469) pp. 36–50.

Burr, Irving W. 1942. Cumulative frequency functions. *The Annals of Mathematical Statistics* **13**(2) pp. 215–232.

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.

Campbell, Patricia, Meriam Dennie, Karen Dougherty, Oksana Iwaskiw, Karen Rollo. 2004. Implementation of an ed protocol for pain management at triage at a busy level i trauma center. *Journal of emergency nursing: JEN: official publication of the Emergency Department Nurses Association* **30**(5) 431.

Carr, Brendan G, Judd E Hollander, William G Baxt, Elizabeth M Datner, Jesse M Pines. 2010. Trends in boarding of admitted patients in us emergency departments 2003–2005. *The Journal of emergency medicine* **39**(4) 506–511.

Carr, Brendan G, Adam J Kaye, Douglas J Wiebe, Vicente H Gracias, C William Schwab, Patrick M Reilly. 2007a. Emergency department length of stay: a major risk factor for pneumonia in intubated blunt trauma patients. *The Journal of Trauma and Acute Care Surgery* **63**(1) 9–12.

Carr, Brendan G, Adam J Kaye, Douglas J Wiebe, Vicente H Gracias, C William Schwab, Patrick M Reilly. 2007b. Emergency department length of stay: a major risk factor for pneumonia in intubated blunt trauma patients. *The Journal of Trauma and Acute Care Surgery* **63**(1) 9–12.

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.

Chalfin, Donald B, Stephen Trzeciak, Antonios Likourezos, Brigitte M Baumann, R Phillip Dellinger, et al. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit*. *Critical care medicine* **35**(6) 1477–1483.

Chan, C. W., Galit Yom-Tov, Gabriel Escobar. 2012. When to use speedup: An examination of service systems with returns. *Working Paper* .

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.

Choi, YF, TW Wong, CC Lau. 2006. Triage rapid initial assessment by doctor (triad) improves

waiting time and processing time of the emergency department. *Emergency medicine journal* **23**(4) 262–265.

Cooper, Julie J, Elizabeth M Datner, Jesse M Pines. 2008. Effect of an automated chest radiograph at triage protocol on time to antibiotics in patients admitted with pneumonia. *The American journal of emergency medicine* **26**(3) 264–269.

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.

Deo, Sarang, Gad Allon, Wuqin Lin. 2013. The impact of hospital size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Working Paper* Forthcoming in "Operations Research".

Deo, Sarang, Itai Gurvich. 2011. Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science* **57**(7) 1300–1319.

Derlet, Robert W, John R Richards, Richard L Kravitz. 2001. Frequent overcrowding in us emergency departments. *Academic Emergency Medicine* **8**(2) 151–155.

deVericourt, Francis, Otis B. Jennings. 2011. Nurse staffing in medical units: A queueing perspective. *Operations Research* **59**(6) 1320–1331.

Dobson, Gregory, Tolga Tezcan, Vera Tilson. 2012. Optimal workflow decisions for investigators in systems with interruptions. *Working Paper* Forthcoming in "Management Science".

Falvo, Thomas, Lance Grove, Ruth Stachura, David Vega, Rose Stike, Melissa Schlenker, William Zirkin. 2007. The opportunity loss of boarding admitted patients in the emergency department. *Academic Emergency Medicine* **14**(4) 332–337.

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.

Gans, Noah, Ger Koole, Avishai Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141.

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.

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.

Gino, Francesca, Gary Pisano. 2008. Toward a theory of behavioral operations. *Manufacturing & Service Operations Management* **10**(4) 676–691.

Goldstein, Jacob. 2008. Is keeping patients waiting in the ER a good business move? URL `http://www.slate.com/id/2195851/`.

Government Accountability Office. 2009. Hospital emergency departments: Crowding continues to occur, and some patients wait longer than recommended time frames. URL `http://www.gao.gov/products/GAO-09-347`.

Green, Linda V. 2006. *Patient Flow: Reducing Delay in Healthcare Delivery*, *International Series in Operations Research & Management Science*, vol. 91, chap. Queuing Analysis in Healthcare. Springer.

Green, Linda V., Sergei Savin, Nicos Savva. 2012. "Nursevendor problem": Personnel staffing int he presence of endogenous absenteeism. *Working Paper* .

Green, Linda V., Sergei Savin, Ben Wang. 2006a. Managing patient service in a diagnostic medical facility. *Operations Research* **54**(1) 11–25.

Green, Linda V, Joao Soares, James F Giglio, Robert A Green. 2006b. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13**(1) 61–68.

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

Gross, D, JF Shortle, J Thompson, CM Harris. 2008. Fundamentals of queueing theory.

Guo, Pengfei, Paul Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Science* **53**(6) 962–970.

Hair, J. F. Jr., R. E. Anderson, R. L. Tatham, W. C. Black. 1995. *Multivaraite Data Analysis*. 3rd ed. Macmillan, New York.

Handel, Daniel A, Joshua A Hilton, Michael J Ward, Elaine Rabin, Frank L Zwemer Jr, Jesse M Pines. 2010. Emergency department throughput, crowding, and financial outcomes for hospitals. *Academic Emergency Medicine* **17**(8) 840–847.

Handel, Daniel A, K John McConnell. 2009. The financial impact of ambulance diversion on inpatient hospital revenues and profits. *Academic emergency medicine* **16**(1) 29–33.

Hasija, Sameer, Edieal J Pinker, Robert A Shumsky. 2005. Staffing and routing in a two-tier call centre. *International Journal of Operational Research* **1**(1) 8–29.

Hassin, R., M. Haviv. 2003. *To queue or not to queue: Equilibrium behavior in queueing systems*, vol. 59. Springer.

Haviv, Moshe, Ya'acov Ritov. 2001. Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions. *Queueing Systems* **38** 495–508.

Henneman, Philip L, Michael Lemanski, Howard A Smithline, Andrew Tomaszewski, Janice A Mayforth. 2009. Emergency department admissions are more profitable than non-emergency department admissions. *Annals of Emergency Medicine* **53**(2) 249.

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

Hillier, Frederick S, Gerald J. Lieberman. 2010. *Introduction To Operations Research*. 9th ed. McGraw-Hill Education.

Hobbs, D., S.C. Kunzman, D. Tandberg, D. Sklar. 2000. Hospital factors associated with emergency center patients leaving without being seen. *The American journal of emergency medicine* **18**(7) 767–772.

Hoot, Nathan R, Dominik Aronsky, et al. 2008. Systematic review of emergency department crowding: causes, effects, and solutions. *Annals of emergency medicine* **52**(2) 126–136.

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

Hsia, R.Y., S.M. Asch, R.E. Weiss, D. Zingmond, L.J. Liang, W. Han, H. McCreath, B.C. Sun. 2011. Hospital determinants of emergency department left without being seen rates. *Annals of emergency medicine* **58**(1) 24.

Hu, Bin, Saif Benjaafar. 2009. Partitioning of servers in queueing systems during rush hour. *Manufacturing & Service Operations Management* **11**(3) 416–428.

Huang, Tingliang, Gad Allon, Achal Bassamboo. 2012. Bounded rationality in service systems. *Working Paper* .

Hui, Michael K., David K. Tse. 1996. What to tell consumers in waits of different lengths: An integrative model of service evaluation. *Journal of Marketing* **60**(2) pp. 81–90.

Hwang, Ula, Melissa L. McCarthy, Dominik Aronsky, Brent Asplin, Peter W. Crane, Catherine K. Craven, Stephen K. Epstein, Christopher Fee, Daniel A. Handel, Jesse M. Pines, Niels K. Rathlev, Robert W. Schafermeyer, Jr Zwemer Frank L., Steven L. Bernstein. 2011. Measures of crowding in the emergency department: A systematic review. *Academic Emergency Medicine* **18**(5) 527–538.

Ibrahim, Rouba, Ward Whitt. 2011. Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations Research* **59**(5) 1106–1118.

Imai, K., I Nonaka, H. Takeuchi. 1985. Managing the new product development process: How japanese compaies learn and unlearn. K. B. Clark, R. H. Hayes, C. Lorenz, eds., *The Uneasy Alliance: Managing the Productivity-Technology Dilemma*. Havard Business School Press, Cambridge, MA.

Institute of Medicine. 2007. Hospital-based emergency care: At the breaking point.

Jackson, James R. 1963. Jobshop-like queueing systems. *Management Science* **10**(1) 131–142.

Jaeker, Jillian B., Anita L. Tucker. 2012. Hurry up and wait: Differential impacts of congestion, bottleneck pressure, and predictabiltiy, on patient length of stay. *Working Paper* .

Janakiraman, N., R.J. Meyer, S.J. Hoch. 2011. The psychology of decisions to abandon waits for service. *Journal of Marketing Research* **48**(6) 970–984.

Jang, Ji Yeon, Sang Do Shin, Eui Jung Lee, Chang Bae Park, Kyoung Jun Song, Adam J. Singer. 2013. Use of a comprehensive metabolic panel point-of-care test to reduce length of stay in the emergency department: A randomized controlled trial. *Annals of Emergency Medicine* **61**(2) 145 – 151.

Jouini, Oualid, Zeynep Aksin, Yves Dallery. 2011. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management* **13**(4) 534–548.

Jouini, Oualid, Yves Dallery, Zeynep Aksin. 2009. Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *International Journal of Production Economics* **120**(2) 389 – 399.

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. 2012. Does multitasking improve performance? Evidence from the emergency department. *Working Paper* .

Kc, Diwas Singh, Christian Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* .

King, Diane L, David I Ben-Tovim, Jane Bassham. 2006. Redesigning emergency department patient flows: application of lean thinking to health care. *Emergency Medicine Australasia* **18**(4) 391–397.

Kremer, Mirko, Laurens Debo. 2012. Herding in a queue: A laboratory experiment. *Working Paper* .

Kulstad, Erik B, Rishi Sikka, Rolla T Sweis, Ken M Kelley, Kathleen H Rzechula. 2010. Ed overcrowding is associated with an increased frequency of medication errors. *The American journal of emergency medicine* **28**(3) 304–309.

Larson, Richard C. 1987. Perspectives on queues: Social justice and the psychology of queueing. *Operations Research* **35**(6) 895–905.

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

Le, MM, FL Zwemer, VJ Dickerson, S Paris. 2004. Providing mobile phones to emergency medicine residents: perceived effects on physician communication and work. *Annals of Emergency Medicine* **44**(4) S28.

Lee, Donald K. K., Stefanos A. Zenios. 2009. Optimal capacity overbooking for the regular treatment of chronic conditions. *Operations Research* **57**(4) 852–865.

Liu, Shan W, Stephen H Thomas, James A Gordon, Azita G Hamedani, Joel S Weissman. 2009. A pilot study examining undesirable events among emergency department–boarded patients awaiting inpatient beds. *Annals of emergency medicine* **54**(3) 381–385.

Loch, Christoph H., Christian Terwiesch. 1998. Communication and uncertainty in concurrent engineering. *Management Science* **44**(8) 1032–1048.

Lu, Yina, Marcelo Olivares, Andres Musalem, Ariel Schilkrut. 2012. Measuring the effect of queues on customer purchases. *Working Paper* .

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.

Maister, David H. 1985. The psychology of waiting lines.

Mandelbaum, A., S. Zeltyn. 2013. Data-stories about (im)patient customers in tele-queues. *Working Paper* .

Mandelbaum, Avishai, Petar Momcilovic. 2012. Queues with many servers and impatient customers. *Mathematics of Operations Research* **37**(1) 41–65.

Mandelbaum, Avishai, Petar Momcilovic, Yulia Tseytlin. 2012. On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. *Management Science* **58**(7) 1273–1291.

Mandelbaum, Avishai, Nahum Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems* **36** 141–173.

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.

McHugh, Megan, Marsha Regenstein, Bruce Siegel. 2008. The profitability of medicare admissions based on source of admission. *Academic Emergency Medicine* **15**(10) 900–907.

Meisel, Zachary, Jesse Pines. 2008. Waiting doom: How hospitals are killing E.R. patients.

Meislin, Harvey W, Sally A Coates, Janine Cyr, Terry Valenzuela. 1988. Fast track: Urgent care within a teaching hospital emergency department: Can it work? *Annals of emergency medicine* **17**(5) 453–456.

Mitka, Mike. 2008. Economics may play role in crowding, boarding in emergency departments. *JAMA: the journal of the American Medical Association* **300**(23) 2714–2715.

Nagler, Jonathan. 1994. Scobit: An alternative estimator to logit and probit. *American Journal of Political Science* **38**(1) pp. 230–255.

Nash, Kathleen, Brian Zachariah, Jennifer Nitschmann, Benjamin Psencik. 2007. Evaluation of the fast track unit of a university emergency department. *Journal of Emergency Nursing* **33**(1) 14–20.

Niska, Richard, Farida Bhuiya, Jianmin Xu, et al. 2010. National hospital ambulatory medical care survey: 2007 emergency department summary. *Natl Health Stat Report* **26**(26) 1–31.

O'Brien, Debra, Aled Williams, Kerrianne Blondell, George A Jelinek. 2006. Impact of streaming. *Australian Health Review* **30**(4) 525–532.

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

Pashler, Harold. 1994. Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin* **116**(2) 220–224.

Pham, J.C., G.K. Ho, P.M. Hill, M.L. McCarthy, P.J. Pronovost. 2009. National study of patient, visit, and hospital characteristics associated with leaving an emergency department without being seen: predicting LWBS. *Academic Emergency Medicine* **16**(10) 949–955.

Pines, Jesse M, John D Heckman. 2009. Emergency department boarding and profit maximization for high-capacity hospitals: Challenging conventional wisdom. *Annals of Emergency Medicine* **53**(2) 256–258.

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, A. Russell Localio, Judd E Hollander, William G Baxt, Hoi Lee, Carolyn Phillips, Joshua P Metlay. 2007. The impact of emergency department crowding measures on time to antibiotics for patients with community-acquired pneumonia. *Annals of emergency medicine* **50**(5) 510–516.

Pines, Jesse M, Charles V Pollack, Deborah B Diercks, Anna Marie Chang, Frances S Shofer, Judd E Hollander. 2009. The association between emergency department crowding and adverse cardiovascular outcomes in patients with chest pain. *Academic Emergency Medicine* **16**(7) 617–625.

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.

Plambeck, Erica, Qiong Wang. 2012. Hyperbolic discounter and queue-length information management for unpleasant services that generate future benefits. *Working Paper* .

Polevoi, Steven K., James V. Quinn, Nathan R. Kramer. 2005. Factors associated with patients who leave without being seen. *Academic Emergency Medicine* **12**(3) 232–236.

Rabin, Elaine, Keith Kocher, Mark McClelland, Jesse Pines, Ula Hwang, Niels Rathlev, Brent Asplin, N Seth Trueger, Ellen Weber. 2012. Solutions to emergency department boarding and crowding are underused and may need to be legislated. *Health Affairs* **31**(8) 1757–1766.

Rask, Kimberly J, Mark V Williams, Ruth M Parker, Sally E McNagny. 1994. Obstacles predicting lack of a regular provider and delays in seeking care for patients at an urban public hospital. *JAMA: the journal of the American Medical Association* **271**(24) 1931–1933.

Roberts, Rebecca R, Paul W Frutos, Ginevra G Ciavarella, Leon M Gussow, Edward K Mensah, Linda M Kampe, Helen E Straus, Gnanaraj Joseph, Robert J Rydman. 1999. Distribution of variable vs fixed costs of hospital care. *JAMA: the journal of the American Medical Association* **281**(7) 644–649.

Rogers, Tessa, Nicola Ross, Daniel Spooner. 2004. Evaluation of a "see and treat" pilot study introduced to an emergency department. *Accident and emergency nursing* **12**(1) 24–27.

Rogg, Jonathan G., Benjamin A. White, Paul D. Biddinger, Yuchiao Chang, David F. M. Brown. 2013. A long-term analysis of physician triage screening in the emergency department. *Academic Emergency Medicine* **20**(4) 374–380.

Rowe, Brian H., Peter Channan, Michael Bullard, Sandra Blitz, L. Duncan Saunders, Rhonda J. Rosychuk, Harris Lari, William R. Craig, Brian R. Holroyd. 2006. Characteristics of patients who leave emergency departments without being seen. *Academic Emergency Medicine* **13**(8) 848–852.

Saghafian, S, W.J. Hopp, M.P. VanOyen, J.S. Desmond, S.J. Kronick. 2013. Complexity-based triage: A tool for improving patient safety and operational efficiency. *Working Paper* .

Saghafian, Soroush, Wallace J. Hopp, Mark P. Van Oyen, Jeffrey S. Desmond, Steven L. Kronick. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* **60**(5) 1080–1097.

Schull, Michael J, Kate Lazier, Marian Vermeulen, Shawn Mawhinney, Laurie J Morrison, et al. 2003a. Emergency department contributors to ambulance diversion: a quantitative analysis. *Ann Emerg Med* **41**(4) 467–476.

Schull, Michael J, Laurie J Morrison, Marian Vermeulen, Donald A Redelmeier. 2003b. Emer-

gency department gridlock and out-of-hospital delays for cardiac patients. *Academic emergency medicine* **10**(7) 709–716.

Schull, Michael J., Marian Vermeulen, Graham Slaughter, Laurie Morrison, Paul Daly. 2004. Emergency department crowding and thrombolysis delays in acute myocardial infarction. *Annals of Emergency Medicine* **44**(6) 577–585.

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.

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.

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

Shiber, Joseph R. 2010. Emergency department admissions and inpatient discharges: A complex relationship. *Annals of Emergency Medicine* **56**(2) 202–203.

Shimkin, Nahum, Avishai Mandelbaum. 2004. Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Systems* **47** 117–146.

Shumsky, Robert A, Edieal J Pinker. 2003. Gatekeepers and referrals in services. *Management Science* **49**(7) 839–856.

Singer, Adam J., Joshua Ardise, Janet Gulla, Julie Cangro. 2005. Point-of-care testing reduces length of stay in emergency department chest pain patients. *Annals of Emergency Medicine* **45**(6) 587 – 591.

Singer, Adam J, Henry C Thode Jr, Peter Viccellio, Jesse M Pines. 2011. The association between length of emergency department boarding and mortality. *Academic Emergency Medicine* **18**(12) 1324–1329.

Solberg, Leif I, Brent R Asplin, Robin M Weinick, David J Magid. 2003. Emergency department crowding: consensus development of potential measures. *Annals of emergency medicine* **42**(6) 824–834.

Soremekun, O., F.S. Shofer, E. Datner, J. Moore, K. Heidi, D. Grasso. 2012. The impact of an emergency department mid-track on patient flow. *Annals of Emergency Medicine* **60**(4, Supplement) S106 –.

Soremekun, Olanrewaju A, Paul D Biddinger, Benjamin A White, Julia R Sinclair, Yuchiao Chang, Sarah B Carignan, David FM Brown. 2011. Operational and financial impact of physician screening in the ED. *The American journal of emergency medicine* .

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

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

Storm-Versloot, Marja N., Dirk T. Ubbink, Johan Kappelhof, Jan S. K. Luitse. 2011. Comparison of an informally structured triage system, the emergency severity index, and the manchester triage system to distinguish patient priority in the emergency department. *Academic Emergency Medicine* **18**(8) 822–829.

Takakuwa, Kevin M, Frances S Shofer, Stephanie B Abbuhl. 2007. Strategies for dealing with emergency department overcrowding: a one-year study on how bedside registration affects patient throughput times. *The Journal of emergency medicine* **32**(4) 337–342.

Takeuchi, Hirotaka, Ikujiro Nonaka. 1986. The new new product development game. *Harvard Business Review* **64**(1) 137 – 146.

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

Thompson, Steven, Manuel Nunez, Robert Garfinkel, Matthew D. Dean. 2009. Efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operations Research* **57**(2) 261–273.

Urgent Matter Learning Network II. 2010. Improving patient flow & reducing emergency department crowding. URL `http://urgentmatters.org/media/file/UM%20LN%20II%20IB%20-%20FINAL%20CORRECTED%202.pdf`. Issue Brief 1.

Walsh, B, WK Yamarick. 2005. Beam me up, Scotty. A new emergency department in Ohio goes live with a wearable, push-button communication system on opening day, reducing noise, improving staff communication and increasing patient privacy. *Health Management Technology* **26**(7) 24–26.

Weber, Ellen J., Ian McAlpine, Barbara Grimes. 2011. Mandatory triage does not identify high-acuity patients within recommended time frames. *Annals of Emergency Medicine* **58**(2) 137 – 142.

Whitt, W. 1984. The amount of overtaking in a network of queues. *Networks* **14** 411–426.

Whitt, Ward. 1999a. Partitioning customers into service groups. *Management Science* **45**(11) 1579–1592.

Whitt, Ward. 1999b. Predicting queueing delays. *Management Science* **45**(6) 870–888.

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.

Wiler, Jennifer L, Christopher Gentle, James M Halfpenny, Alan Heins, Abhi Mehrotra, Michael G Mikhail, Diana Fite. 2010. Optimizing emergency department front-end operations. *Annals of emergency medicine* **55**(2) 142–160.

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.

Yankovic, Natalia, Linda V. Green. 2012. A queuing model for nurse staffing. *Working Paper* .

Zhang, Zhe George, Hsing Paul Luh, Chia-Hung Wang. 2011. Modeling security-check queues. *Management Science* **57**(11) 1979–1995.

Zohar, Ety, Avishai Mandelbaum, Nahum Shimkin. 2002. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science* **48**(4) 566–583.