# Waiting Patiently: An Empirical Study of Queue Abandonment in an Emergency Department

Robert J. Batt, Christian Terwiesch
The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, batt@wharton.upenn.edu, terwiesch@wharton.upenn.edu

We study queue abandonment from a hospital emergency department. 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. For example, observing an additional person in the queue or an additional arrival to the queue leads to an increase in abandonment probability equivalent to a fifteen minute or nine minute increase in wait time respectively. We also 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 customer response to visual queue elements is not currently accounted for in most queuing models. Additionally, 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 waiting customers.

*Key words*: Healthcare operations; Service Operations; Empirical; Queues with Abandonment
*History*: Submitted to Management Science; April, 2013

## 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. 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.

## 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. 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 1999, 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.

## 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 9).

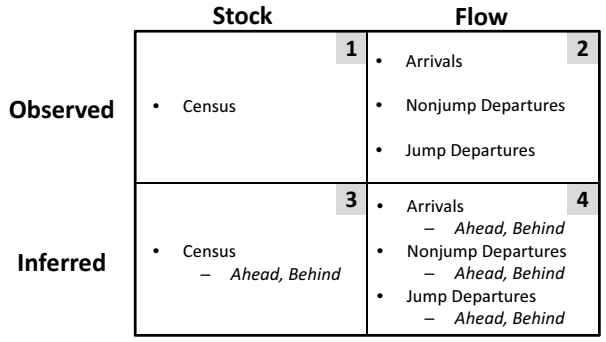
We now develop a theory of how patients respond to visible queue elements. Abstracting from the optimal stopping problem formulation of Aksin et al. (2012), we assume that the abandonment decision is the result of a patient repeatedly evaluating the following personal utility function:

$$\text{Utility} = \max\left[\mathbb{E}\left[\begin{pmatrix}\text{Service}\\ \text{Reward}\end{pmatrix} - \begin{pmatrix}\text{Wait}\\ \text{Cost}\end{pmatrix} \times \begin{pmatrix}\text{Residual}\\ \text{Wait}\end{pmatrix}\right], 0\right] \tag{1}$$

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

**Figure 1** **Visible Queue State Variables**




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 observes. 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 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.

## 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 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 7.1 we follow Zohar et al. (2002) and take averages across time to estimate the system waiting time. In Section 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 ±0.3hours, and more than 80% of the values are between ±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

$$OWAIT_i = \begin{cases} WAIT_i & \text{if patient stays} \\ \widehat{WAIT}_i & \text{if patient abandons} \end{cases} \tag{2}$$

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.

## 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 ($\varepsilon$) that represents the difference in the unobserved random component of the utility of each option. Since $\varepsilon$ 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 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

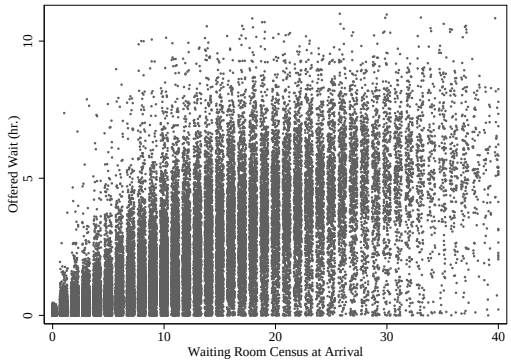
$$\begin{aligned}\text{Prob}\left(LWBS_i = 1|\mathbf{x}\right) =& \Phi(\beta_0 + \beta_1 OWAIT_i + \beta_2 CENSUS_i + \beta_3 OWAIT_i \times CENSUS_i \\ &+ \beta_4 TRITEST_i + \mathbf{X_i}\boldsymbol{\beta_P} + \mathbf{Z_i}\boldsymbol{\beta_T})\end{aligned} \tag{3}$$

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. 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 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.

**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.*

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 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 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, 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:

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

$$\begin{aligned} LWBS_i^* =& \beta_{2,0} + \beta_{2,1} OWAIT_i + \beta_{2,2} CENSUS_i + \beta_{2,3} OWAIT_i \times CENSUS_i \\ & + \beta_{2,4} TRITEST_i + \mathbf{X_i}\boldsymbol{\beta_{2,P}} + \mathbf{Z_i}\boldsymbol{\beta_{T2,T}} + \varepsilon_{2,i} \\ & LWBS_i = 1 \text{ if } LWBS_i^* > 0,\ 0 \text{ otherwise} \end{aligned} \tag{5}$$

$\mathbf{X_i}$ and $\boldsymbol{Z_i}$ are specified as before in Equation 3. $\varepsilon_1$ and $\varepsilon_2$ are assumed to be standard bivariate normally distributed with correlation coefficient $\rho$. 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.

## 7. Results

### 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

**Figure 3** Pr(LWBS) vs. Wait Time

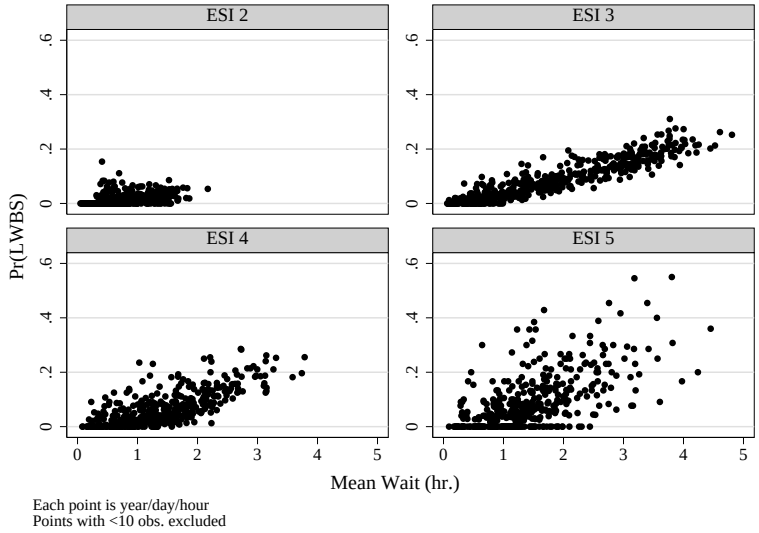



**Table 2** Model Fit Measures of Regressing Pr(LWBS) on Wait Time

| | 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 |

suggests that customers become *more* patient when the system is busy. We find no such evidence in the ED.

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.

### 7.2. Regression Analysis

The graphs in Section 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 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 7.1. For clarity, we focus on results for triage level ESI 3 in Subsections 7.2.1 and 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 7.2.3, and in Subsection 7.2.4 we examine the impact of triage testing on all triage levels.

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

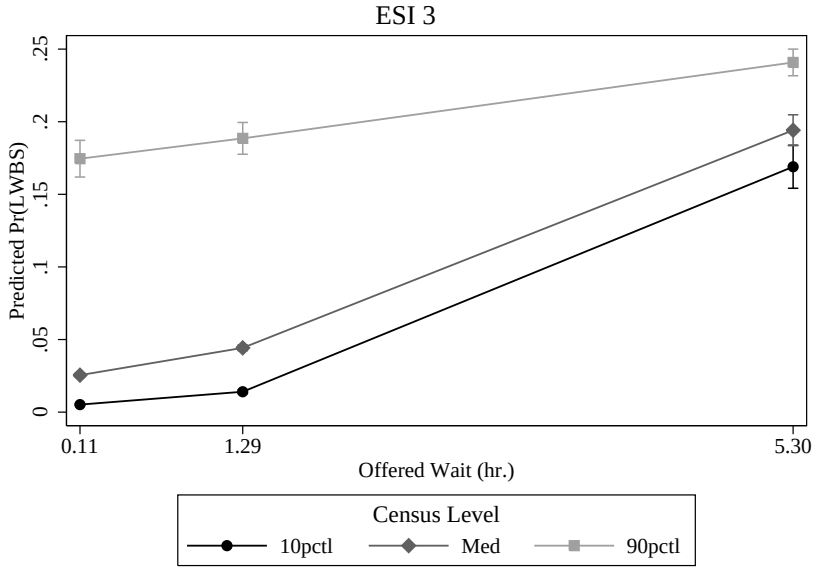
| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Offered Wait (hr.) | 0.20*** | 0.12*** | 0.11*** | 0.11*** |
| | (0.00) | (0.01) | (0.01) | (0.01) |
| Census | 0.07*** | 0.06*** | 0.06*** | 0.06*** |
| | (0.00) | (0.00) | (0.00) | (0.00) |
| Wait x Census | -0.01*** | -0.01*** | -0.01*** | -0.01*** |
| | (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
$^{*}$ $p<0.10$ , $^{**}$ $p<0.05$ , $^{***}$ $p<0.01$

**7.2.1. Observed Variables** Model 1 of Table 3 shows the results of estimating Equation 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

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




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 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 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.

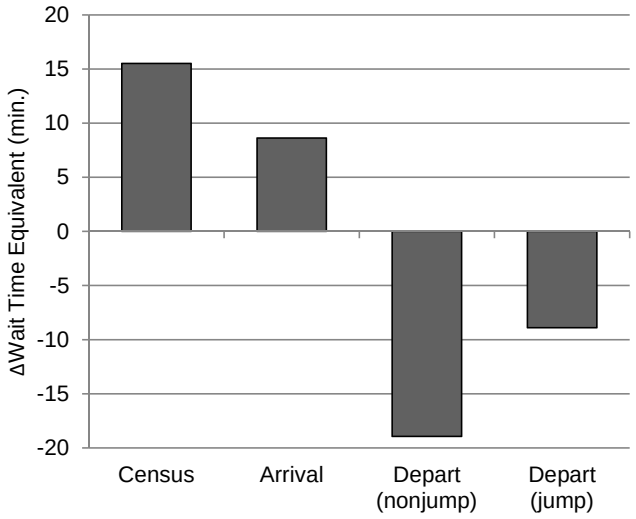
The large gap between the median and 90th percentile census levels even for very short waits suggests that large crowds lead to rapid abandonment even when the actual wait time is low. This also explains why the slope of the 90th percentile census line is relatively flatter. People are likely abandoning 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) 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

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




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

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.

**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 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 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 each ahead/behind pair and find that the values are significantly different and that the ahead component

**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 |

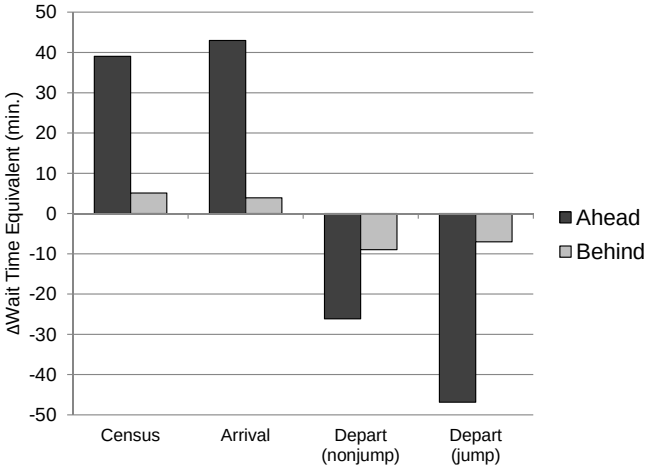
Cluster robust standard errors in parentheses
Controls not shown: Age, Gender, Insurance, Pain, Triage Test, Year, Weekend, Block of Day
* $p<0.10$ , ** $p<0.05$ , *** $p<0.01$

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

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




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

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

$$\theta = \frac{\beta_{AHEAD}}{\beta_{AHEAD} + \beta_{BEHIND}} \tag{6}$$

Let $\beta_{AHEAD}$ be the estimated coefficient of one of the Ahead variables in Table 4 and let $\beta_{BEHIND}$ 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 $\beta_{AHEAD}$would be non-zero, $\beta_{BEHIND}$ would be zero and $\theta$ would be unity. If, however, patients had no ability to discern those ahead and behind, then $\beta_{AHEAD}$ would equal $\beta_{BEHIND}$ and $\theta$ 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 $\theta$ range between 0.75 and 1. While we do not interpret $\theta$ 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.

**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

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

| | (1) ESI 2 | (2) ESI 3 | (3) ESI 4 | (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
* $p<0.10$ , ** $p<0.05$ , *** $p<0.01$

more sensitive to 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.

**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) |
| $\rho$ | | | | | -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
* $p < 0.10$ , ** $p < 0.05$ , *** $p < 0.01$

**7.2.4. Triage Testing** Models 1 through 4 of Table 6 show the results of estimating the basic probit model of Equation 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 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 5 using a bivariate probit model. For ESI 3 and ESI 4 patients, the estimated correlation coefficient ($\rho$) is negative and significant indicating correlation in the error terms of Equations 4 and 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 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.

## 8. Robustness of Model Selection

As mentioned in Section 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 $\varepsilon$ 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 $\varepsilon$ 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 $\varepsilon$ is distributed with cumulative distribution function $F(\boldsymbol{x\beta}, \alpha) = 1 - 1/\left\{1 + \exp\left(\boldsymbol{x\beta}\right)\right\}^{\alpha}$. 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

**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

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

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 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 7.2, underestimates the marginal effect of offered wait and census relative to the better fitting models. Thus, the results presented are conservative.

## 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.

## References

ACEP. 2012. Publishing wait times for emergency department care Http://www.acep.org/clinical—practice-management/publishing-wait-times-for-emergency-department-care,-june-2012.

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

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

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

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

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* .

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.

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.

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.

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

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

Chan, Carri W., Mor Armony, Nicholas Bambos. 2011. Fairness in overloaded parallel queues. *Working Paper* .

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

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.

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

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.

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.

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.

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.

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.

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

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.

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 Ak?in. 2009. Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *International Journal of Production Economics* **120**(2) 389 – 399, <ce:title>Special Issue on Introduction to Design and Analysis of Production Systems</ce:title>.

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

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

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

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

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

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.

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.

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

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

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

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.