

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 patient abandonment from a hospital emergency department queue. We find that patients are influenced by what they see around them in the waiting room: waiting room census, arrivals, and departures. Patients are also sensitive to being "overtaken" in the line. Lastly, patients also appear to make inferences about the severity of the patients around them and respond differently to people more sick and less sick moving through the system. The fact that patients respond to visual stimulus suggests that in order to reduce patient abandonment hospitals should actively manage what patients see and what information they have regarding the wait.

Key words: Healthcare operations; Service Operations; Empirical; Queues with Abandonment

History: Working Paper; December, 2012

1. Introduction

The body of knowledge on queuing theory is voluminous and spans more than half 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 renegeing) 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 (ED), abandonment takes on the added dimension of the risk of a patient suffering an adverse medical event. While the hospital may or 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 (Larson 1987). 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, we observe many service systems, such as call centers and emergency

departments, which provide limited or no information to waiting customers. Thus, it is an open question as to which is better: a hidden queue, a fully visible queue, or some middle-ground, semi-visible queue. This is an active area of analytical queuing theory research (e.g. Guo and Zipkin 2007, Armony et al. 2009), but there is limited empirical work on the subject.

We examine this question in the setting of a hospital emergency department. Most EDs are naturally semi-visible in that 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 required service time of others is not readily apparent. Interestingly, most 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 leaving who need care (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 specifically on how what the patient observes during the waiting phase of the queuing encounter impacts the abandonment decision. We perform a detailed econometric study of patient queuing behavior in an ED. Using data from the electronic patient tracking system, we are able to identify factors both before and during the queuing encounter which impact the abandonment decision. We make the following three contributions:

1. We show that observed queue length has an effect on abandonment separate from its direct effect on wait time.
2. We show that the observed flow of patients in and out of the waiting room has an effect on abandonment. Furthermore, we show that patients respond differently to outflows that maintain priority based first-come-first-served order and those that do not.
3. We show that patients respond to more than just the “facts” that they observe. They make assumptions about the severity of other patients and respond differently to the flow of more and less severe patients.

Taken together, these contributions show that patient abandonment behavior is affected by what the waiting patients observe. However, since the patients do not have full knowledge of the system state or design, they may respond in ways that bias them toward excessive abandonment.

2. Clinical Setting

Our study is based on data from a large, urban, teaching hospital with an average of 4,700 ED visits per month over the study period of January, 2009 through December, 2011. The ED has 25 treatment rooms and 15 hallway beds for a theoretical maximum treatment capacity of 40 beds.

However, the actual treatment capacity at any given moment can fluctuate for various reasons. The hospital also operates an express lane or FastTrack (FT) for low acuity patients. The FT is generally open from 8am to 8pm on weekdays, and from 9am to 6pm on weekends. The FT operates somewhat autonomously from the rest of the ED in that it utilizes seven dedicated beds and is usually staffed by dedicated group of Certified Registered Nurse Practitioners (CRNP) rather than Medical Doctors (MD)¹.

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. 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 acuity, using the five-level Emergency Severity Index (ESI) triage scale with 1 being most severe and 5 being least severe (Gilboy et al. 2011). Patients are generally not informed of their assigned triage level. The triage nurse also has the option of ordering some diagnostic tests, for example an x-ray or a blood test.

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 cannot be sure which people in the waiting room are patients (versus family and friends), or 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.

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.

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

Most patients likely have little or no understanding that the ED and FT exist 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 other 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).

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 he is willing to wait, and he waits in the queue until he either enters service or reaches his maximum wait time, at which point he 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 Garnett et al. (2002), 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 three terms of interest in these models: the reward for service, the instantaneous unit waiting cost, and the residual waiting time (Shimkin and Mandelbaum 2004).²

There is a rich literature of studies which use such utility-based models. For example, Mandelbaum and Shimkin (2000) considers customer abandonment from a system with a “fault state” in which service will never be initiated. Customers continuously update their expected residual wait time and eventually conclude that they are likely in the fault state and thus abandon. Likewise, in Hassin and Haviv (1995) the reward from service may drop to zero thereby inducing abandonment. See Hassin and Haviv (2003) for a review of various assumptions that lead to rational abandonments.

² Some models also include a discount rate, which adds another term of interest.

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) show 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.

Related questions of what to tell waiting customers and when to tell them have 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 Whitt (1999), Ibrahim and Whitt (2009, 2011b,a). 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" question under the assumption of strategic behavior by both customers and providers. Allon and Bassamboo (2011) shows that providers can benefit from delaying information announcements because doing so allows the provider more time to observe the state of the system.

There are many empirical studies from fields such as marketing and behavioral studies which 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).

Two recent papers in the Operations Management literature study customer queue behavior empirically. 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 the assumptions of an invisible queue and linear waiting costs, 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.

Lu et al. (2012) examines how elements of a visible queue, such as queue length and number of servers, effect 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 do not appropriately incorporate all available information into their balk or abandon decisions.

Our work differs from these two related works in several ways. First, our setting is different in that we examine a semi-visible queue; in the ED, the waiting room is visible but the service area is not. Further, patients do not know the characteristics of the other waiting patients. Thus, patients have access to some information about the queue, but it may be of limited value. Second, because we have more granular than in Lu et al. (2012), we can allow for more factors, such as the flow of other patients (customers), to influence the abandonment decision. Lastly, in terms of methodology, we use reduced form models since we are not estimating any latent structural parameters as in Aksin et al. (2012). We hope our work will continue to expand the understanding of customer behavior while waiting in line.

4. Framework & Hypotheses

The underlying operative question that we explore is, “Should hospitals provide queue-status information to waiting patients?” As mentioned in Section 1, the default stance in many EDs is to provide no information. Patients are not informed of such things as their own triage level, the triage level of others, the queue length, the expected wait time, or whether or not there is a FastTrack. Two reasons for this behavior are that providing accurate information, especially wait time predictions, can be difficult, and that patients may respond to perceived bad news by abandoning. This logic is in line with a key finding in the “queues with information” literature; when providers and customers have different objectives, there can be incentives to hide the queue. For example, in Guo and Zipkin (2007), the provider maximizes throughput and the customer maximizes personal utility. Under this assumption, the provider may have an incentive to hide the queue when the queue length is longer than the customers’ uninformed expectation of queue length. Hiding the queue effectively tricks customers, who would otherwise abandon if they had full information, into staying. Or, as mentioned in Section 3, Plambeck and Wang (2012) shows that hiding the queue can be welfare-maximizing even if it does not maximize customers’ short-term utility. In essence, hiding the queue fools customers into doing what is best for themselves in the long run, even if they do not recognize it at the time.

This literature highlights the need to define each party's objective. For the patient, we assume a desire to maximize personal utility. As described in Section 3, the utility function is comprised of three terms: the reward for service, the instantaneous unit waiting cost, and the residual waiting time. For the hospital, the objective is less clear. Revenue maximization suggests serving everyone who walks in the door. Likewise, a belief in a moral obligation to serve all comers leads to a desire to eliminate abandonment. Welfare maximization suggests providing full information if the hospital believes that patients can accurately evaluate their own utility. However, if the hospital believes that patients can not accurately assess their need for treatment, then the hospital may withhold information. Finally, profit maximization suggests 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. Therefore, for our study we assume that the hospital is seeking to minimize abandonment.

With the objective defined as minimizing abandonment of utility maximizing patients, the question of interest becomes, "Does the current policy of providing no queue-status information serve to minimize abandonment?" However, to be clear, just because the hospital does not provide queue status information does not mean that the patients are completely in the dark. The ED is *not* an invisible queue. Recall from Section 2 that the patients all wait in a single waiting room and are able to observe what goes on around them. If they choose to, patients can be aware of the number of people in the waiting room and the flow of patients in and out of the waiting room. Thus, we are interested in whether patients are responding to this visual stimulus and if doing so leads to higher or lower abandonment. Understanding the impact of these visual cues on abandonment will help identify possible ways to modify the information available to patients. Our hope is that this empirical study will provide the justification necessary to receive approval from the Institution Review Board for a controlled experiment in the ED.

We focus on the impact of four categories of variables created by the permutations of two pairs of conditions: stocks and flows, and observed and inferred. The key "stock" of interest is the waiting room census, while the key "flows" are the arrivals and departures from the waiting room. By observed and inferred we mean that some things can be objectively observed, such as the number of

arrivals to the ED, while others can only be inferred, such as the number of patients in the waiting room with a higher triage classification than one's own.

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, as described earlier, then the waiting room census likely impacts the patient's residual time estimate and abandonment behavior just as in Guo and Zipkin (2007) and Plambeck and Wang (2012). This leads to our first hypothesis.

HYPOTHESIS 1. Abandonment increases with waiting room census.

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 increase their residual time estimate upward leading to more abandonment.

HYPOTHESIS 2. Abandonment increases with observed arrival rate.

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 the observer. It signals system speed, which presumably reduces the residual time estimate. However, the jump departure does not move the observer any closer to service, and thus the reduction in residual time estimate is less than for a regular departure. Further, the observer may view the jump as unfair and be more likely to abandon. These possibilities lead to the following two hypotheses.

HYPOTHESIS 3. Abandonment decreases with observed departure rate.

HYPOTHESIS 4. Jump departures decrease abandonment less than regular departures.

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 may impact their wait time differently than less sick patients. However, since all patient information is kept confidential, patients are left to infer relative acuteness by simply observing the other waiting patients. Certainly, this is an inexact process at best, but likely not a pointless endeavor.

If patients are inferring relative acuteness of other patients, then this leads to a much more complex set of hypotheses that essentially splits each of the above hypotheses into two parts, one for more acute and one for less acute patients. For example, arrivals of more sick patients will likely increase abandonment since waiting patients may fear that the new arrivals will jump them in line for service. In contrast, arrivals of less sick patients may have no impact on abandonment since they presumably will be served later than the current waiting patients. The presence of the FastTrack, and most patient's lack of awareness of such, further complicates the picture. For example, an ESI 4 patient may observe the arrival of a much sicker looking ESI 2 patient and be tempted to abandon thinking that his own waiting time just got longer. However, if the FastTrack is open, that ESI 4 patient is likely to be served in the FastTrack and the arrival of an ESI 2 patient will have no impact on his wait. So, a knowledgeable patient would react one way to an arrival and an ignorant patient would react another. Rather than enumerate all the potential responses to the inferred stock and flow variables, we simply state the following general hypothesis and discuss the specifics in Section 7.2.

HYPOTHESIS 5. Abandonment behavior is affected differently by relatively higher and lower acuity patients.

5. Data Description & Definitions

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 level³.

³We do not include ESI 1 patients because these patients never abandon. However, we include ESI 1 patients in all relevant census measures.

Table 1 Summary Statistics

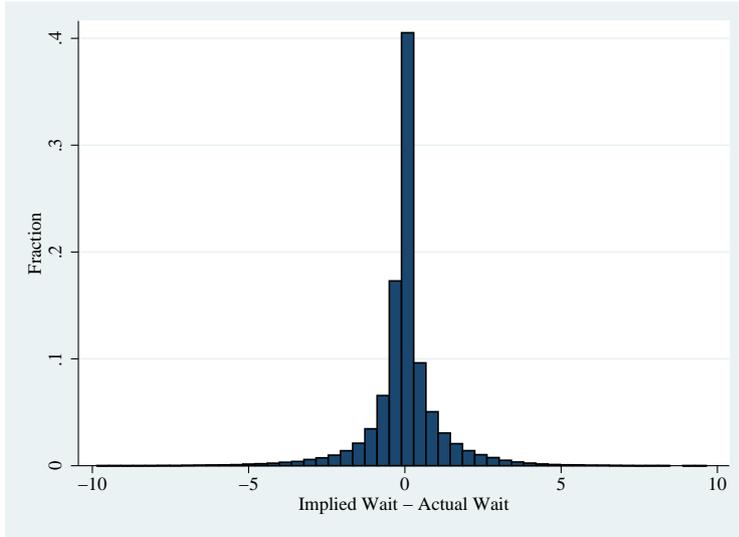
	ESI 2	ESI 3	ESI 4	ESI 5
Age	49.8 (0.11)	39.0 (0.07)	34.7 (0.07)	34.2 (0.14)
%Female	54% (0.003)	66% (0.002)	58% (0.002)	51% (0.005)
Pain (1-10)	4.5 (0.03)	5.5 (0.02)	5.4 (0.02)	4.1 (0.04)
%FastTrack	2% (0.001)	3% (0.001)	68% (0.002)	67% (0.005)
Wait Time(hr.)	1.0 (0.01)	1.9 (0.01)	1.3 (0.01)	1.3 (0.01)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

Means shown. Standard error of mean in parentheses

Performing empirical analysis on customer abandonment is inherently challenging due to the censored or missing nature of the data. Ideally, one would observe each customer’s maximum willingness to wait and the actual wait time if he stayed. But, only the minimum of these two is ever realized, leading to censored data. Further, in our data, actual 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 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. For patients who did remain, we calculate this directly from the timestamps. For patients who abandon, we must estimate the offered wait with what we refer to as $i \pm 1$ estimation. We sort the patients into chronological order of arrival within each triage class. Then for each abandoned patient, we calculate the average of the wait times of the two chronologically adjacent patients (one before and one after) who did not abandon. More specifically, we define the variable $WAIT_i$ as the observed wait time for patient i of triage level T_i . For patients that abandoned, this value is missing (NA). We then calculate “carry forward” and “carry backward” variables as

$$WAIT_{-cf_i} = \begin{cases} WAIT_i & \text{if } WAIT_i \neq NA \\ WAIT_{-cf_{i-1}} & \text{if } WAIT_i = NA \end{cases} \quad (1)$$

Figure 1 Histogram of Accuracy of Imputed Wait Time

$$WAIT_cb_i = \begin{cases} WAIT_i & \text{if } WAIT_i \neq NA \\ WAIT_cb_{i+1} & \text{if } WAIT_i = NA \end{cases} \quad (2)$$

From these we calculate the imputed wait time for all patients as

$$\widehat{WAIT}_i = \frac{1}{2} (WAIT_cf_i + WAIT_cb_i) \quad (3)$$

To get a sense of the accuracy of the imputed wait time, we examine the deviation between \widehat{WAIT}_i and $WAIT_i$ for all patients that did not abandon. Figure 1 shows a histogram of this difference across all patients for whom a wait time is observed. The deviation has a mean of 0.00 and a standard deviation of 1.1 hours. 50% of the values are between ± 0.3 hours, and more than 80% of the values are between ± 1 hour. Thus, the imputed wait appears to be unbiased, and is relatively close to the true value.

We then define the offered wait time as follows

$$OWAIT_i = \begin{cases} WAIT_i & \text{if } WAIT_i \neq NA \\ \widehat{WAIT}_i & \text{if } WAIT_i = NA \end{cases} \quad (4)$$

Another key independent variable of interest is the waiting room census. To calculate this 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 $CENSUS_t$ 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 ($CENSUS_{t,T}$, $T \in \{1, 2, 3, 4, 5\}$) for later use. We assign a census value to each patient ($LOAD_i$) based on the time of arrival. For example, for patient i who arrives at time interval t , $LOAD_i = CENSUS_t$.

In order to test Hypothesis 5, we would ideally decompose $LOAD_i$ into those patients whom patient i perceives to be more sick and less sick than herself. However, since these perceptions are not observed, 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, and any waiting patient of lower priority (higher ESI number) is considered as behind the arriving patient. These definitions are expressed as follows for patient i of ESI triage level T_i who arrived at time t .

$$LOAD_AHEAD_i = \sum_{j=1}^{T_i} CENSUS_{t,j} \quad (5)$$

$$LOAD_BEHIND_i = \sum_{j=T_i+1}^5 CENSUS_{t,j} \quad (6)$$

The flow variables are constructed from the patient timestamps. For each patient visit we calculate the number of arrivals ($ARRIVE_i$), nonjump departures ($NONJUMP_i$), and jump departures ($JUMP_i$) that occur within one hour of patient i 's arrival. As with the census variable, we also decompose the flow variables by triage level ($ARRIVE_{i,T}$, $NONJUMP_{i,T}$, $JUMP_{i,T}$, $T \in \{1, 2, 3, 4, 5\}$). We also split each flow variable into two parts based on those ahead and behind the given patient based on the triage level-based priority queuing. For arrivals, only those in higher priority classes are considered ahead of patient i since they should be the only patients that have the potential to jump ahead of patient i .

$$ARRIVE_AHEAD_i = \sum_{j=1}^{T_i-1} ARRIVE_{i,j} \quad (7)$$

$$ARRIVE_BEHIND_i = \sum_{j=T_i}^5 ARRIVE_{i,j} \quad (8)$$

For nonjump departures, patients of the same triage classification or higher are classified as ahead of patient i since these patients should be served before patient i according to the assumed priority queuing discipline. Patients of lower triage score are classified as behind patient i .

$$NONJUMP_AHEAD_i = \sum_{j=1}^{T_i} NONJUMP_{i,j} \quad (9)$$

$$NONJUMP_BEHIND_i = \sum_{j=T_i+1}^5 NONJUMP_{i,j} \quad (10)$$

Lastly, jump departures of higher triage class than patient i are classified as ahead since the priority queuing discipline would have them be served first. Patients of equal or lower triage classification are classified as behind patient i .

$$JUMP_AHEAD_i = \sum_{j=1}^{T_i-1} JUMP_{i,j} \quad (11)$$

$$JUMP_BEHIND_i = \sum_{j=T_i}^5 JUMP_{i,j} \quad (12)$$

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 ($\mathbf{x}\boldsymbol{\beta}$) plus a random variable (ε) that represents the difference in the unobserved random component of the utility of each option. Since ε is stochastic, these models can only predict a probability of choosing one action over the other. The choice of distribution of ε determines the functional form of the response of the prediction to a change in an independent variable. Choosing either the logistic or the normal distribution leads to the well known logit and probit models, respectively. Assuming ε follows a complementary log log distribution ($F(\mathbf{x}\boldsymbol{\beta}) = 1 - \exp[-\exp(\mathbf{x}\boldsymbol{\beta})]$) leads to the cloglog model. The Burr-10 distribution (Burr 1942) assumes ε is distributed with cumulative distribution function $F(\mathbf{x}\boldsymbol{\beta}, \alpha) = 1 - 1/\{1 + \exp(\mathbf{x}\boldsymbol{\beta})\}^\alpha$. As a regression model, it is referred to as the skewed logistic or scobit model. Note that the logit model is a special case of the scobit model with $\alpha = 1$.

Selecting the best model a priori is difficult because each has theoretical or practical advantages and disadvantages. 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. However, the logit and probit models are symmetric about $\mathbf{x}\boldsymbol{\beta} = 0$, 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, cloglog coded two ways, and scobit coded two ways.

We fit all models and find that indeed the the asymmetric models generally provide the best fit based on the Bayesian Information Criterion. 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. Therefore, for the body of the paper we present the results from the logit model because the results have a direct odds-ratio interpretation and because the reader is likely most familiar with this type of model. See Section 8 for comparison of the other models.

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

$$\begin{aligned} \text{logit} [\Pr (LWBS_i)] = & \beta_0 + \beta_1 OWAIT_i + \beta_2 LOAD_i + \beta_3 OWAIT_i \times LOAD_i \\ & + \mathbf{X}_i \boldsymbol{\beta}_P + \mathbf{Z}_i \boldsymbol{\beta}_T \end{aligned} \quad (13)$$

\mathbf{X}_i is a vector of patient-visit specific covariates including age, gender, insurance type, chief complaint, pain level, and a dummy variable indicating if any diagnostics are ordered at triage. \mathbf{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. We estimate the model separately for each triage level between 2 and 5.

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 served patients, the offered wait is not known until service begins, at which point $LWBS$ 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 flipping a mental coin 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 his total probability of abandoning. Thus, the $OWAIT$ variable picks up this effect, that patients who are given the opportunity to be in the system longer are more likely to abandon, even though the actual offered wait value is not observed by the patient.

Our identification strategy is based on the assumption that $OWAIT$ and $LOAD$ are not perfectly correlated and both contain some amount of 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, auxilliary 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.

One potential concern with this model specification is the collinearity between $OWAIT$ and $LOAD$. In fact, the pairwise correlation between $OWAIT$ and $LOAD$ is roughly 0.72, which is high enough to be of concern. However, the Variance Inflation Factors (VIF) for the model in Equation 13

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 census variables used in all models. When we do this for Equation 13, the VIFs now range from 2.4 to 3.2, well within the acceptable range of collinearity.

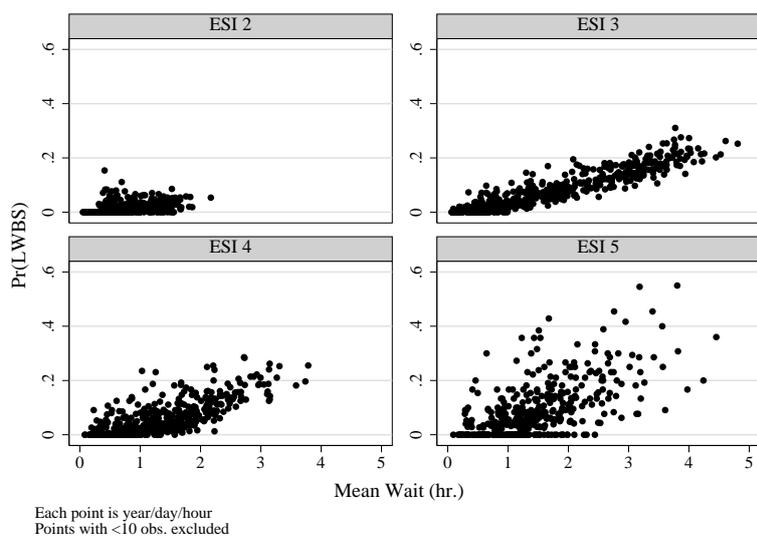
As we examine each of the hypotheses, we gradually add more variables to the model of Equation 13. For instance, we replace $LOAD_i$ with the pair of variables $LOAD_AHEAD_i$ and $LOAD_BEHIND_i$. We do this both without and then with the interaction of the load variables with the $OWAIT_i$ variable.

Once we add the flow variables to the model, we must restrict the sample. If we include all patients in the flow analysis we will get results that are difficult to interpret because a patient who spends only a short time in the waiting room may observe very few arrivals or departures simply because she was not exposed to the system for very long. In contrast, a patient who is in the waiting room for several hours has much more opportunity to observe the system and be influenced by the flows. To account for this, we restrict the sample to only patients with an offered wait of greater than one hour. Recall from Section 5 that the flow variables ($ARRIVE_i$, $NONJUMP_i$, $JUMP_i$, etc.) are defined as the flows during the first hour after arrival of patient i . Thus, 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.

Another challenge of the sample restriction is that we do not know when abandoning patients abandon. We restrict the sample to patients with an offered time of greater than one hour, but 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 patients abandon quickly before observing many arrivals or departures, this should bias our results toward the null hypothesis of no effect. Thus, any significant results are likely conservative estimates of the impact of the flow variables.

Because some patients in our data have multiple visits to the ED during the study period, the data could be considered unbalanced panel data and analyzed using binary panel data methods. However, since about 40% of the patients have only a single visit, models such as the fixed-effects logit would only be estimable for the patients with multiple visits and who sometimes stay and sometimes abandon. Therefore, rather than use panel methods 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 observations for a single individual. It also adjusts for potential misspecification of the functional form of the model.

Figure 2 Pr(LWBS) vs. Wait Time



7. Results

7.1. Overview Graphs

It is informative to begin by using scatter plots to visualize the relationship between abandonment and wait time, following the example of Zohar et al. (2002). 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 2 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 ($3 \text{ years} \times 7 \text{ days} \times 24 \text{ hours} = 504$). 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 plot of Figure 2 is for a given 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 2. First, there is an increasing linear trend for all triage levels (Table 2). This is different from Zohar et al. (2002), in that Zohar et al. (2002) finds the surprising result that the probability of abandonment does not increase with expected wait (the linear fit is flat). This suggests that customers become *more* patient when the system is busy. We find no such evidence in the ED. We point out our “unsurprising” result here because we refer back to it and build upon it later.

The second feature we observe in Figure 2 is that the dispersion from the linear trend increases with decreasing patient acuteness. Table 2 quantifies this effect by the root mean squared error (RMSE) for linear regressions for each of the graphs in Figure 2. Further, from the R^2 values in

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

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 that wait time does not capture. ESI 2 appears somewhat different. While ESI 2 displays a positive linear trend with little dispersion (significant positive slope and low RMSE in Table 2), the model has the lowest R^2 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 (an unintended result).

Given that wait time only partially explains the observed abandonment behavior, we now turn to logistic regression models to better understand the operational drivers of abandonment and the differences across triage classes.

7.2. Regression Analysis

The graphs in Section 7.1 were based on means calculated by aggregating across year/day/hour combinations. We now drill down a level and use the logistic 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. 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 at the end of the section.

Table 3 shows the results of estimating Equation 13 (model 3), as well as two simpler models (models 1 and 2) and two more complex models (models 4 and 5). All of these models are without flow variables and thus are estimated on the full sample.

Model 1 establishes the expected baseline result that abandonment increases with offered wait just as the Erlang-A model suggests. The logit coefficient can be directly interpreted as the change

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

	(1)	(2)	(3)	(4)	(5)
Offered Wait	0.33***	0.21***	0.37***	0.20***	0.36***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Load		0.06***	0.14***		
		(0.00)	(0.00)		
Wait x Load			-0.02***		
			(0.00)		
Load(Ahead)				0.07***	0.17***
				(0.00)	(0.00)
Load(Behind)				0.02***	0.05***
				(0.00)	(0.01)
WaitxLoad(Ahead)					-0.03***
					(0.00)
WaitxLoad(Behind)					-0.01***
					(0.00)
N	65,622	65,622	65,622	65,622	65,622
McFadden's R^2	0.17	0.19	0.21	0.19	0.21
BIC	34,513	33,792	32,890	33,723	32,766

Cluster robust standard errors in parentheses

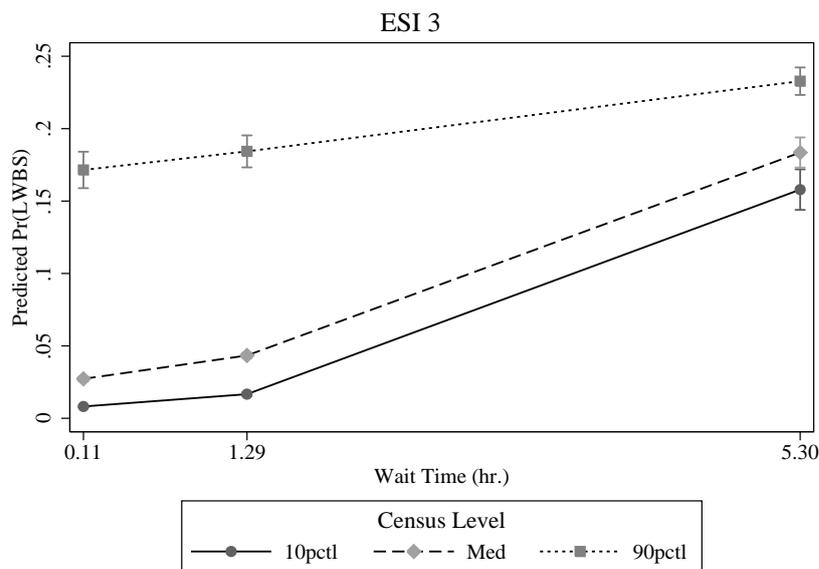
Controls not shown: Age, Gender, Insurance, Pain, Year, Weekend×Block of Day

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

in the log odds of abandonment with a change in the independent variable, or the exponentiated coefficient is factor of change of the odds ratio with a change in the independent variable. Thus, Model 1 shows that the log odds of abandoning increase by 0.33 if the offered wait increases by 1 hour. Likewise, the odds ratio of abandonment increases by 39% ($\exp(0.33) = 1.39$) if the offered wait goes up by 1 hour. Note that this suggests that patients are not all abandoning immediately, for if they were the offered wait coefficient would be insignificant.

In Model 2, we add in the load or census variable as an explanatory variable. Both offered wait and load are positive and significant which supports Hypothesis 1, that the census level increases abandonment. This 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. While it appears that Offered Wait has a larger impact on abandonment, we must be cognizant of the scaling of the explanatory variables. Offered wait is in units of hours with a mean of 2.1 and a standard deviation of 2.1, and Load is in units of people with a mean of 11.7 and a standard deviation of 9.0. Thus a one standard deviation change in Offered Wait leads to a .41 increase in log odds of abandonment while a one standard deviation increase in Load leads to a 0.6 increase in log odds of abandonment.

Model 3 adds in the interaction of Offered Wait and Census. The increased McFadden's R^2 and decreased Bayesian Information Criterion (BIC) indicate that Model 3 is a better fit than Models 1 & 2. The first-order terms of Offered Wait and Load are remain positive and significant ($\beta_1, \beta_2 > 0$),

Figure 3 Predicted Pr(LWBS) as a function of Offered Wait and Load

but the negative interaction coefficient makes interpreting marginal effects more difficult. Predicted values will be more informative.

Figure 3 shows the predicted abandonment probabilities at three levels of wait time and wait 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 Load level. The three lines are the 10th, 50th, and 90th percentile Load levels (1, 10, and 25 people). 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 load 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 almost 17%. Thus, Figure 3 shows that patients respond to both increasing wait time and waiting room census with increased abandonment.

The large gap between the median and 90th percentile census points 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. Many people are abandoning sooner and are not sticking around to be impacted by the experienced wait. In contrast, for low to mid crowding, the effect of long wait times becomes much larger as the wait times approach several hours.

Models 4 and 5 of Table 3 begin to explore Hypothesis 5 by using the variables *LOAD_AHEAD* and *LOAD_BEHIND* in place of the single *LOAD* variable used in the first three models. Model 4 shows that the presence of “sicker” people has more than three times the impact on abandonment as does the presence of “healthier” people. Recall, that patients are not told which waiting patients are ahead or behind in line, so the fact that the coefficients for Load(Ahead) and Load(Behind) are significantly different is strong evidence that patients are able to visually infer the relative status of those in the waiting room. Model 5 includes the interactions of Offered Wait and the load variables. While this model provides the best fit (lowest BIC), like Model 3 it is difficult to directly interpret the coefficients. We again turn to predicted values.

Table 4 Predicted Pr(LWBS) for ESI 3 Patients (Offered Load fixed at mean)

		Load(Behind)		
		1	5	10
Load(Ahead)	1	0.03	0.03	0.04
	5	0.05	0.05	0.06
	10	0.08	0.08	0.09
	15	0.12	0.14	0.15
	20	0.20	0.21	0.24

Table 4 shows the predicted abandonment probability as a function of the load ahead and behind. Just as with Model 4, we see that the marginal effect of a person ahead is larger than the marginal effect of a person behind. This concurs with intuition that those who are less sick and are thus behind in line have less impact on the behavior of a given patient. The fact that Load(Behind) has any impact is likely due to the imperfect nature of inferring the relative status of those in the waiting room.

To examine Hypotheses 2, 3, and 4, 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 1 of Table 5 is the same as Model 3 from Table 3 but with the restricted sample. Note that the coefficients in the two models are identical in their sign and somewhat similar in magnitude, which suggests that the patient behavior of the restricted sample is similar to that of the full sample.

Model 2 adds in variables for the number of arrivals and the total number of departures into service (regardless of jump or nonjump status). The positive and significant coefficient on Arrivals supports Hypothesis 2 that arrivals lead to more abandonments. The negative and significant coefficient on Departures supports Hypothesis 3 that observing departures leads to reduced abandonment. Figure 4 plots predicted abandonment probabilities for 10th, 50th, and 90th percentile values of both arrivals and departures and gives a sense of the magnitude of the impact of these variables.

Table 5 Effect of Wait Time, Census, and Flow on Pr(LWBS) [ESI 3]

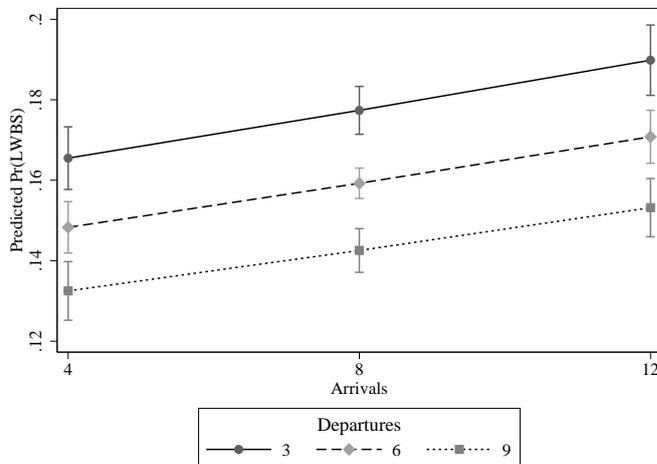
	(1)	(2)	(3)
Offered Wait	0.21*** (0.01)	0.20*** (0.01)	0.20*** (0.01)
Load	0.11*** (0.00)	0.11*** (0.00)	0.12*** (0.00)
Wait x Load	-0.02*** (0.00)	-0.02*** (0.00)	-0.02*** (0.00)
Arrivals		0.02*** (0.01)	0.02*** (0.01)
Departures		-0.05*** (0.01)	
Departures(nonjump)			-0.05*** (0.01)
Departures(jump)			-0.02 (0.02)
N	35,855	35,855	35,855
McFadden's R^2	0.10	0.10	0.10
BIC	28,782	28,727	28,735

Cluster robust standard errors in parentheses

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

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

Figure 4 Effect of Arrivals and Departures of Pr(LWBS)



Model 3 of Table 5 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 supports both Hypothesis 3 and Hypothesis 4. The negative coefficient on nonjump departures shows that waiting patients view these departures as a good sign of processing speed and progress towards service, thus people are less likely to abandon. In contrast, the insignif-

icant 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.

Table 6 Effect of Ahead/Behind variables on Pr(LWBS) [ESI 3]

	(1)	(2)	(3)
Offered Wait	0.21*** (0.01)	0.19*** (0.01)	0.19*** (0.01)
Load(Ahead)	0.14*** (0.01)	0.14*** (0.01)	0.14*** (0.01)
Load(Behind)	0.02 (0.01)	0.02* (0.01)	0.02* (0.01)
WaitxLoad(Ahead)	-0.02*** (0.00)	-0.02*** (0.00)	-0.02*** (0.00)
WaitxLoad(Behind)	-0.00 (0.00)	-0.00 (0.00)	-0.00 (0.00)
Arrival(Ahead)		0.08*** (0.01)	0.09*** (0.01)
Arrival(Behind)		0.01 (0.01)	0.01 (0.01)
Depart(Ahead)		-0.06*** (0.01)	
Depart(Behind)		-0.02* (0.01)	
Depart(Nonjump-Ahead)			-0.05*** (0.01)
Depart(Nonjump-Behind)			-0.02* (0.01)
Depart(Jump-Ahead)			-0.10*** (0.03)
Depart(Jump-Behind)			-0.01 (0.02)
N	35,855	35,855	35,855
McFadden's R^2	0.10	0.11	0.11
BIC	28,671	28,607	28,625

Cluster robust standard errors in parentheses

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

Year, Weekend, Block of Day

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

Table 6 parallels Table 5 with the change that each stock and flow variable is now split into its ahead and behind components. We continue to restrict the sample to those with an offered wait of greater than one hour. Model 1 of Table 6 is the same as Model 5 of Table 13 but with the restricted sample. Not surprisingly given the large drop in sample size, we find fewer coefficients to be significant in the restricted model. Those coefficients that are significant (Offered Wait, Load(Ahead), WaitXLoad(Ahead)) maintain the same sign and similar magnitudes. Comparing the models in Table 6 with the parallel models in Table 5 we note that the models in Table 6 with variables split

by Ahead/Behind all show a lower BIC indicating a better fit despite the fact that several variables are not significant. Thus, this further supports Hypothesis 5 that patients respond differently to patients they perceive as ahead or behind them in line. More specifically, we see in all the models in Table 6 that it is the “Ahead” variables that have the largest impact on abandonment behavior, which is consistent with rational behavior ignoring those behind them in line. Again, it is the fact that patients are able to detect priority line order with a reasonable amount of accuracy that is perhaps most interesting.

Table 7 Effect of Ahead/Behind variables on Pr(LWBS)

	(1)	(2)	(3)	(4)
	ESI 2	ESI 3	ESI 4	ESI 5
Offered Wait	0.31*** (0.10)	0.19*** (0.01)	0.28*** (0.03)	-0.01 (0.05)
Load(Ahead)	0.31*** (0.04)	0.14*** (0.01)	0.07*** (0.01)	0.08*** (0.01)
Load(Behind)	0.06*** (0.02)	0.02* (0.01)	0.16*** (0.05)	
WaitxLoad(Ahead)	-0.04*** (0.01)	-0.02*** (0.00)	-0.01*** (0.00)	-0.00 (0.00)
WaitxLoad(Behind)	-0.01 (0.01)	-0.00 (0.00)	-0.01 (0.01)	
Arrival(Ahead)	0.24 (0.26)	0.08*** (0.01)	0.05*** (0.01)	0.05*** (0.02)
Arrival(Behind)	0.01 (0.02)	0.01 (0.01)	0.02 (0.02)	-0.04 (0.06)
Depart(Ahead)	-0.16*** (0.05)	-0.06*** (0.01)	-0.05*** (0.01)	-0.05*** (0.02)
Depart(Behind)	-0.04 (0.03)	-0.02* (0.01)	-0.11** (0.05)	0.02 (0.39)
N	8,974	35,855	19,745	5,213
McFadden’s R^2	0.10	0.11	0.14	0.11
BIC	2,673	28,607	9,564	3,581

Cluster robust standard errors in parentheses

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

Year, Weekend, Block of Day

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

We conclude this section with Table 7 showing the results for all triage levels of the best fitting model (Model 2 from Table 6). The results are similar across triage levels in terms of which coefficient are significant and the sign of those coefficients. At first glance, there appear to be two unexpected results in ESI 4 (Model 3). The Load(Behind) coefficient is larger than the Load(Ahead) coefficient, and the Depart(Behind) coefficient is larger than the Depart(Ahead) coefficient. This would seem to suggest that ESI 4 patients are somehow more sensitive to those behind than in front of them. However a Wald test for coefficient equality shows that the two Load coefficient are not significantly

different, 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.

Triage level 5 is the most dissimilar of the four models. Observe that the Offered Wait has an insignificant effect on abandonment while Load(Ahead) continues to lead to greater abandonment.⁴ 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.

8. Robustness Analyses

In this section we provide results of various alternative assumptions model specifications for the sake of establishing the robustness of the presented results.

As mentioned in Section 6, there are several binary outcome models to choose from. Table 8 compares six such model specifications for the baseline model with offered wait, load, and the interaction for ESI 3 (cross-reference Table 3, model 3). The top panel of the table shows estimated

	(1)	(2)	(3)	(4)	(5)	(6)
	logit	probit	cll-lwbs	cll-stay	scobit-lwbs	scobit-stay
<i>Coefficients</i>						
Offered Wait	0.37*** (0.01)	0.20*** (0.00)	0.32*** (0.01)	-0.16*** (0.00)	0.63*** (0.04)	-0.17*** (0.01)
Load	0.14*** (0.00)	0.07*** (0.00)	0.12*** (0.00)	-0.05*** (0.00)	0.21*** (0.01)	-0.06*** (0.00)
Wait x Load	-0.02*** (0.00)	-0.01*** (0.00)	-0.02*** (0.00)	0.01*** (0.00)	-0.03*** (0.00)	0.01*** (0.00)
alpha					0.123 (0.015)	11.2 (5.43)
<i>Marginal Effects</i>						
Offered Wait	0.023*** (0.001)	0.026*** (0.001)	0.021*** (0.001)	-0.030*** (0.001)	0.032*** (0.001)	-0.029*** (0.001)
Load	0.006*** (0.000)	0.006*** (0.000)	0.005*** (0.000)	-0.007*** (0.000)	0.007*** (0.000)	-0.007*** (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

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

⁴ The variable *LOAD_BEHIND* is not included in the ESI 5 because ESI 5 is the lowest priority level.

coefficients for the variables of interest. The middle panel shows marginal effects of the variables of interest at their respective means. The bottom panel gives model fit statistics. We see that all the models are similar in terms of fit as indicated by both the log-likelihood and the BIC. The scobit-lwbs 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 we 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 percent increase in abandonment, while a one unit increase in load leads to a 0.5% to 0.7% increase in abandonment. Note that the logit model, which we used for the presentation of main results in Section 8, underestimates the marginal effect of offered wait and load relative to the better fitting models. Thus, the results we presented are conservative.

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. It is possible that there are unobserved variables, for example pallor, that are common, or at least correlated, to 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.

One way to control for potential correlated omitted variables is 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 probit models with error terms ε_1 and ε_2 respectively. ε_1 and ε_2 are assumed to be standard bivariate normally distributed with correlation coefficient ρ . If $\rho = 0$, this indicates that the control variables are adequately controlling for the endogenous triage testing and the models can be estimated separately without significant bias.

Table 9 compares the results of a regular probit model to a bivariate probit model for ESI 3 and 4. For ESI 3, the estimated correlation coefficient (ρ) is significant indicating correlation in the error terms. Despite this significant correlation, the coefficients of the offered wait and load terms are essentially the same between the models. What does change dramatically is the coefficient on the Triage Test dummy variable. Without controlling for the correlated errors, one would conclude that triage testing leads to a large reduction in abandonment. However, once we control for the correlated

Table 9 Comparing Probit and Bivariate Probit Models

	(1)	(2)	(3)	(4)
	ESI 3	ESI 3	ESI 4	ESI 4
	Probit	Biprobit	Probit	Biprobit
Offered Wait	0.195*** (0.005)	0.192*** (0.005)	0.220*** (0.012)	0.227*** (0.012)
Load	0.072*** (0.002)	0.068*** (0.002)	0.042*** (0.002)	0.041*** (0.002)
Wait x Load	-0.011*** (0.000)	-0.011*** (0.000)	-0.006*** (0.001)	-0.006*** (0.001)
Triage Test (Y/N)	-0.511*** (0.019)	-0.068 (0.056)	-0.466*** (0.036)	-0.348** (0.151)
ρ		-0.27*** (0.031)		-0.10 (0.087)
N	65,622	65,631	39,806	39,806

Cluster robust standard errors in parentheses

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

Chief Complaint, Year, Weekend, Block of Day

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

errors, that effect disappears and we see that triage testing is not a cause of abandonment for these patients. In contrast, for ESI 4 patients, ρ is insignificant indicating that we need not worry about estimating the two equation model. Also note that the coefficients of offered wait, load, and even the triage test dummy are all quite similar between the two ESI 4 models. Thus we conclude that for our purposes of examining the effects of wait and load on abandonment, the single equation model, which is simpler to estimate and work with, is sufficient.

9. Summary & 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. We confirm prior study findings that wait time is a determinant of abandonment. More interestingly, we also find that the queue length (waiting room census, in our study) is predictive of abandonment 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 respond differently to observing exits that maintain versus violate first-come first-served order. Further, waiting patients appear to infer the relative health status of those around them and respond differently to those more sick and less sick. For example, we find that arrival of sicker patients increases abandonment more so than does the arrival of less sick patients. This is presumably because patients recognize that sicker patients will likely be served first.

The essence of our contribution is in providing evidence that waiting customers (patients) glean information from watching the queue around them. While prior work has shown abandonment to

be influenced by such things as playing music and providing distractions, ours is the first to show customers responding to the actual functioning of the queue, to the operational state variables of the system. This is managerial relevant for any organization that wants to actively manage customer abandonment. In the ED, where the goal is minimization of abandonment, our results suggest 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 wait time estimates may serve to reduce abandonment.

Future work should take these findings and use them to motivate and inform a series of controlled experiments. The experiments could focus on how providing additional information modifies the patient response to observed queue behavior. For example, it would be interesting to compare the effectiveness of providing a wait time estimate versus providing the patient's queue position versus providing full queue status. It is not a priori obvious what intervention of information will result in abandonment reduction. Another avenue for experimentation would be to explore how obscuring queue information affects abandonment. Presumably, obscuring the queue would shift the behavior toward the Erlang-A model, but this should be explored empirically.

Acknowledgments

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

References

- ACEP. 2012. Publishing wait times for emergency department care URL <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. 2011. The impact of delaying the delay announcements. *Operations Research* **59**(5) 1198–1210. doi:10.1287/opre.1110.0972. URL <http://or.journal.informs.org/content/59/5/1198.abstract>.
- 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. doi:10.1287/opre.1110.0976. URL <http://or.journal.informs.org/content/59/6/1382.abstract>.
- Armony, Mor, Nahum Shimkin, Ward Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57**(1) 66–81. doi:10.1287/opre.1080.0533.
- Baccelli, F., G. Hebuterne. 1981. On queues with impatient customers. *Performance* 159–179.

- Bitran, Gabriel R., Juan-Carlos Ferrer, Paulo Rocha e Oliveira. 2008. Om forumÜmanaging customer experiences: Perspectives on the temporal aspects of service encounters. *Manufacturing & Service Operations Management* **10**(1) 61–83. doi:10.1287/msom.1060.0147. URL <http://msom.journal.informs.org/content/10/1/61.abstract>.
- 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. 10.1023/A:1015781818360.
- 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. URL <http://www.jstor.org/stable/27590517>.
- Burr, Irving W. 1942. Cumulative frequency functions. *The Annals of Mathematical Statistics* **13**(2) pp. 215–232. URL <http://www.jstor.org/stable/2235756>.
- 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. URL <http://www.gpo.gov/fdsys/pkg/FR-2012-07-30/pdf/2012-16813.pdf>.
- Gans, Noah, Ger Koole, Avishai Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141. doi:10.1287/msom.5.2.79.16071.
- Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4**(3) 208–227. doi:10.1287/msom.4.3.208.7753. URL <http://msom.journal.informs.org/content/4/3/208.abstract>.
- Gilboy, N, T Tanabe, D Travers, AM Rosenau. 2011. *Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Implementation Handbook*. Agency for Healthcare Research and Quality, Rockville, MD, 4th ed. AHRQ Publication No. 12-0014.
- Greene, William H. 2012. *Econometric Analysis*. 7th ed. Prentice Hall.
- Guo, Pengfei, Paul Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Science* **53**(6) 962–970. doi:10.1287/mnsc.1060.0686. URL <http://mansci.journal.informs.org/content/53/6/962.abstract>.
- Hair, J. F. Jr., R. E. Anderson, R. L. Tatham, W. C. Black. 1995. *Multivariate Data Analysis*. 3rd ed. Macmillan, New York.
- Hassin, R., M. Haviv. 1995. Equilibrium strategies for queues with impatient customers. *Operations Research Letters* **17**(1) 41–45.
- Hassin, R., M. Haviv. 2003. *To queue or not to queue: Equilibrium behavior in queueing systems*, vol. 59. Springer.

- 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.
- Ibrahim, Rouba, Ward Whitt. 2009. Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management* **11**(3) 397–415. doi:10.1287/msom.1080.0223.
- Ibrahim, Rouba, Ward Whitt. 2011a. Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals. *Productions and Operations Management* **20**(5) 654.
- Ibrahim, Rouba, Ward Whitt. 2011b. Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations Research* **59**(5) 1106–1118. doi:10.1287/opre.1110.0974.
- 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. doi:10.1287/msom.1110.0339. URL <http://msom.journal.informs.org/content/13/4/534.abstract>.
- 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. doi:10.1016/j.ijpe.2008.01.011. URL <http://www.sciencedirect.com/science/article/pii/S0925527309000140>. <ce:title>Special Issue on Introduction to Design and Analysis of Production Systems</ce:title>.
- Larson, Richard C. 1987. Or forumÜperspectives on queues: Social justice and the psychology of queueing. *Operations Research* **35**(6) 895–905. doi:10.1287/opre.35.6.895.
- Lu, Yina, Marcelo Olivares, Andres Musalem, Ariel Schilkrot. 2012. Measuring the effect of queues on customer purchases. *Working Paper* .
- Mandelbaum, Avishai, Petar Momcilovic. 2012. Queues with many servers and impatient customers. *Mathematics of Operations Research* **37**(1) 41–65. doi:10.1287/moor.1110.0530. URL <http://mor.journal.informs.org/content/37/1/41.abstract>.
- Mandelbaum, Avishai, Nahum Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems* **36** 141–173. URL <http://dx.doi.org/10.1023/A:1019131203242>. 10.1023/A:1019131203242.
- Nagler, Jonathan. 1994. Scobit: An alternative estimator to logit and probit. *American Journal of Political Science* **38**(1) pp. 230–255. URL <http://www.jstor.org/stable/2111343>.
- 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. doi:10.1197/j.aem.2004.10.029. URL <http://dx.doi.org/10.1197/j.aem.2004.10.029>.
- Shimkin, Nahum, Avishai Mandelbaum. 2004. Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Systems* **47** 117–146. URL <http://dx.doi.org/10.1023/B:QUES.0000032804.57988.f3>.
- Whitt, Ward. 1999. Predicting queueing delays. *Management Science* **45**(6) 870–888. doi:10.1287/mnsc.45.6.870.
- 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. doi:10.1287/mnsc.48.4.566.211.