Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Robert J. Batt, Christian Terwiesch

To cite this article:
Robert J. Batt, Christian Terwiesch (2015) Waiting Patiently: An Empirical Study of Queue Abandonment in an Emergency Department. Management Science 61(1):39-59. http://dx.doi.org/10.1287/mnsc.2014.2058

MANAGEMENT SCIENCE
Vol. 61, No. 1, January 2015, pp. 39–59
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2014.2058

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

Robert J. Batt
Wisconsin School of Business, University of Wisconsin–Madison, Madison, Wisconsin 53706, rbatt@bus.wisc.edu

Christian Terwiesch
The Wharton School and Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, terwiesch@wharton.upenn.edu

We study queue abandonment from a hospital emergency department. We show that abandonment is influenced by the queue length and the observable queue flows during the waiting exposure, even after controlling for wait time. 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 25-minute or 5-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.

Keywords: healthcare operations; service operations; empirical; queues with abandonment
History: Received April 6, 2013; accepted August 8, 2014, by Yossi Aviv, operations management.

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. Although 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, provide limited or no information to waiting customers. One reason for this is that uninformed customers might naïvely estimate the waiting time to be short and thus join a queue that they would not otherwise 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 2013). 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. (2013), 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 arrival, 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 patients. The position of the American College of Emergency Physicians (2012) is that providing queue

information might have "unintended consequences" and lead to patients who need care leaving without treatment. 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 time-stamp 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 have rationally 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 who 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, whereas others have to be inferred. For example, the number of patients in the waiting room is directly observable to the patient, whereas, given that the priority data are 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 to predicting and estimating abandonment behavior of ED patients allows us to make the following three observations:

1. We find that for patients of moderate severity, observing an additional patient in the queue increases the probability of abandonment by 1 percentage point, even when appropriately controlling for wait time. This is equivalent to a 25-minute increase in wait time and extends the prior result of Lu et al. (2013) from a deli counter to an emergency department.

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 0.7 percentage points, equivalent to an 18-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" 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 arrival of an additional patient sicker than oneself increases the probability of abandonment by 1 percentage point, whereas observing the arrival of a patient less sick than oneself has no discernible effect on abandonment.

These observations show that patient abandonment behavior is affected by the waiting patients' experience while in the waiting room. Thus, a queue is not simply either visible (such as in a grocery store) or invisible (such as in a call center) but oftentimes combines aspects of both. In such settings, providing no information to customers does not mean that customers are without queue information. Furthermore, to the extent that 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 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" for low-acuity patients. The FastTrack is generally open from 8 A.M. to 8 P.M. on weekdays and from 9 A.M to 6 P.M. on weekends. The FastTrack 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 who 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 2014). 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 the most severe and 5 being the 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. Furthermore, 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. One reason for violation of FCFS is care-provider load balancing. The doctors and nurses working the ED are assigned a block of rooms at the beginning of each shift and are responsible for treating whoever is placed in the block. When a room becomes available, a charge nurse selects which waiting patient to put in the room. The charge nurses do some amount of informal load balancing and therefore might, for example, select an "easy" patient from the waiting room, even if that patient is not next in line, to assign to a doctor or nurse who is working several difficult cases. Also, FCFS is sometimes violated if the charge nurse determines that there is a clinical reason for a person to jump over part of the line.

If the FastTrack is open, then the FastTrack will serve triage level 4 and 5 patients in (roughly) FCFS order by triage level and the ED will serve patients of triage levels 1–3 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 no more acute patients who need immediate attention. Similarly, the FastTrack might serve a triage level 3 patient if he or she has been waiting a long time and the patient's needs can be met by the nurse practitioners in the FastTrack.

Most patients likely have little or no awareness that the ED and FastTrack coexist and work as separate service channels. Furthermore, since patients exit the waiting room through the same doors to receive service in either the ED or the FastTrack, 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. In our study sample, 6.5% of patients abandoned the queue, but this rate varies widely by triage level (see §5). 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 and related variations first introduced by Palm (1943). In the Erlang-A-type models, 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 assumed to be independent and identically distributed draws from the exponential distribution. Later works have introduced variations of the Erlang-A model that relax particular assumptions of the basic model (e.g., Baccelli and Hebuterne 1981, Brandt and Brandt 1999). Examples of work using the Erlang-A model include Brown et al. (2005) and Garnett et al. (2002).

Modeling abandonment in this way provides analytical tractability but does not account for *why* customers abandon or the actual drivers of customer behavior. In short, the Erlang-A model simply assumes that abandonment occurs and proceeds to describe the behavior of the queue under that assumption. Thus, the Erlang-A model is useful for studying how to manage a queue, such as in the call-center staffing literature (e.g., Armony and Mandelbaum 2011, Liu and Whitt 2012), but it is not well suited for

examining how events that occur during the waiting period affect the abandonment decision, which is the focus of this paper.

An alternative view of queue abandonment is based on customer utility maximization. In such models, customers are assumed to be forward-looking, utility-maximizing decision makers who 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, Akşin et al. 2013). Some models also include a discount rate, which adds a fourth term of interest (e.g., Plambeck and Wang 2013).

One of the key findings from this body of literature is that abandoning the queue is not rational in most $M/M/c$-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 that modifies the basic queue model to allow for rational abandonments. For example, nonlinear waiting costs (Haviv and Ritov 2001, Shimkin and Mandelbaum 2004) and a decreasing hazard rate of being served (Mandelbaum and Shimkin 2000) are features that can generate rational queue abandonment in a utility-based queuing framework. Similarly, bounded rationality has been shown to affect queue joining behavior (Kremer and Debo 2013, Huang et al. 2013) and may well impact queue abandonment, but this has not been studied. See Hassin and Haviv (2003) for a review of assumptions that lead to rational abandonments.

The current work is related to this body of literature in that we take a utility framework point of view and examine behavior in a complex queue to examine factors that influence abandonment. Mandelbaum and Shimkin (2000) note that features such as a variable number of servers, variable arrival rates, priorities, incorrect customer beliefs, and real-time information—all features that exist in an ED—can potentially lead to rational queue abandonment. As described in detail in §4, we focus on how real-time observation of the system affects patient abandonment.

There are many studies from a variety of fields that identify drivers of queue abandonment. Although 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) study 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) provide 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. These studies can also be related to patient behavior through the utility framework outlined above. For example, demographic factors such as age, income, and race have all been shown to influence patient abandonment, most likely through different waiting costs or perhaps the benefit or reward of service (Polevoi et al. 2005, Pham et al. 2009). Likewise, institutional factors (e.g., hospital ownership and the presence of medical residents) and operational factors (e.g., utilization level) have also been shown to influence patient abandonment, presumably through systematic differences in the residual waiting times (e.g., hospitals with medical residents have longer service times and longer wait times because both the resident and the attending physician must see the patient) (Hobbs et al. 2000, Polevoi et al. 2005, Hsia et al. 2011). This paper differs from these studies in that it focuses on real-time information available to waiting patients rather than high-level fixed effects.

Although there are several recent empirical operations management papers dealing with queuing systems in the healthcare setting (e.g., Berry Jaeker and Tucker 2012, Batt and Terwiesch 2014, Chan et al. 2014), with the exception of a working paper by Bolandifar et al. (2014), none has focused on queue abandonment. There are, however, two recent papers that study queue abandonment empirically in nonmedical settings, one in a call center and one at a deli.

Akşin et al. (2013) use a structural model to estimate latent 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. This work helps validate the utility-maximization framework of queue abandonment in which customers are balancing a reward from service against the expected cost of waiting.

Our work differs from Akşin et al. (2013) in terms of setting, purpose, and methodology. Our study setting is a *semivisible* multiclass queue (in the ED, the waiting room is visible but the clinical treatment area is not) compared with an *invisible* multiclass queue. This makes our setting more complex because some information becomes available to the customers/patients

during the wait. In terms of purpose and methodology, because the purpose of Akşin et al. is to estimate latent structural parameters, the authors impose strong structural assumptions (known common hazard function, linear waiting costs, etc.). In contrast, our purpose is to identify the impact of observable system events (e.g., arrivals and departures) rather than latent structural parameters, and thus we use reduced-form models that require fewer structural assumptions.

Lu et al. (2013) examine 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 do a poor job of incorporating available information into their balk or abandon decisions.

Our work differs from Lu et al. (2013) in several ways. First, our setting is quite different. Lu et al. examine a fully visible single-class FCFS queue compared with the semivisible multiclass queue in the ED. The additional complexity of the ED setting lets us examine issues related to flow and fairness that are not relevant in simpler settings. Second, because our data are richer and more detailed than in Lu et al., 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. must infer *whether* a customer observed the queue, *when* a customer observed the queue, and the *length* of the queue observed. In contrast, our data allow us to know both when a patient entered the queue and what the queue length was at that moment. Furthermore, we observe the dynamics of the queue during the waiting experience including arrivals, departures, and the patient mix. In terms of results, our work both confirms and contradicts the findings of Lu et al. (2013). We confirm, in both direction and magnitude, their finding regarding customers' response to queue length: longer queues lead to more balking or abandonment. However, in terms of key learnings, our results contradict Lu et al. (2013). Lu et al. show that customers are not very sophisticated in their use of information, even in a very simple queue with access to full information (i.e., they ignore the number of servers). In contrast, our results show that patients in an ED are quite sophisticated both in responding to what they see and in correctly guessing the status of other patients and responding appropriately (e.g., they differentiate between sicker and healthier patients). Thus, our work provides new insights and serves to expand the understanding of the behavior of customers waiting in line.

4. Framework and 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 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 §9.

We now develop a theory of how patients respond to visible queue elements. Abstracting from the optimal stopping problem formulation of Akşin et al. (2013), we assume that the abandonment decision is the result of a patient repeatedly evaluating the following personal utility function:

$$\text{Utility} = \max\big[\mathbb{E}\big[(\text{Service reward}) - (\text{Wait cost}) \times (\text{Residual wait})\big], 0\big]. \quad (1)$$

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. Although 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 because 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. Thus, the following hypothesis development focuses on how an observed event impacts the expected residual wait. Following Akşin et al. (2013), 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: (1) stocks and flows and (2) observed and inferred (see Figure 1). The key "stock" of interest is the waiting room census, and the key "flows" are the arrivals and departures from the waiting room. By "observed" and "inferred," we mean that some

Figure 1 Visible Queue State Variables

	Stock	Flow
Observed	1 • *CENSUS*	2 • *ARRIVALS* • *NONJUMP DEPARTURES* • *JUMP DEPARTURES*
Inferred	3 • *CENSUS* – ahead, behind	4 • *ARRIVALS* – ahead, behind • *NONJUMP DEPARTURES* – ahead, behind • *JUMP DEPARTURES* – ahead, behind

things can be objectively observed, such as the number of arrivals to the ED, whereas 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*. The waiting room census is the first, and perhaps most salient, visual cue that a waiting patient observes. However, both the Erlang-A model and the utility model predict that the waiting room census should not have any impact on abandonment conditional on the patient having entered the system (not balked) and controlling for the wait time. (Anecdotal evidence indicates that very few patients balk upon arrival to the ED.) In the Erlang-A model, the wait time is the only driver of abandonment probability once the customer has entered the system. In the utility model, census should only impact that initial decision whether to balk or enter. Once in the system, only new information that changes the expected residual wait should affect abandonment. However, in contrast to these theoretical predictions, Lu et al. (2013) provides empirical evidence of customers overreacting to the queue length. Given that patients in the ED have less queue status information on which to base their actions than do the deli customers in Lu et al., it is likely that ED patients will also respond with apparent overreaction. This leads to our first hypothesis.

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

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 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 from the waiting room into the service area to begin treatment. Patients who 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 patien A—then this provides a mixed signal to patient A. The departure signals system speed, which should lead to a reduced residual time estimate. However, the jump departure does not move patient A any closer to service, and thus the reduction in residual time estimate should be less than for a regular departure. Furthermore, a jump departure may signal to a waiting patient that she is a low-priority patient who is likely to get jumped again, thus lengthening her residual time estimate.[1] Finally, there may also be a psychological effect on patient A if she 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 nonjump departures.*

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

We frame Hypothesis 2C as a comparison between jump and nonjump departures because we cannot predict a priori the direction of the net effect of jump departures. All we can say is that a jump departure should have a less negative impact on abandonment than would a nonjump departure since a nonjump departure is unambiguously good for the waiting patient. 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. Although 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

[1] We thank an anonymous reviewer for suggesting this reaction.

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 whether patients are able to differentiate between those who are ahead of and behind them in the priority queue and if this affects their behavior. Although we leave the precise definitions of the quadrant 3 and 4 variables to §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 him or her in the queue since those behind the patient should not impact his or her wait time. For example, observing a larger number of patients arriving to the ED of higher priority than a given patient (*ARRIVALS_AHEAD*) should increase abandonment (assuming Hypothesis 2A is true), whereas the number of arrivals of equal or lower priority (*ARRIVALS_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 these hypotheses in terms of comparing the effects of the ahead and behind variables.

Hypothesis 3. *Controlling for wait time, abandonment increases more with the census of those ahead in the priority queue than 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 Hypotheses 3, 4A, 4B, and 4C, 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.

5. Data Description, Definitions, and Study Design

We now describe the data set and define the key variables. In the discussion below, the index t indicates a 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 have multiple visits, and we control for this with clustered standard errors (described in detail in §6). Furthermore, because we estimate all models separately for each triage class, the index i is actually an index within the triage class.

Our data include patient-level information on more than 180,000 patient visits to the ED; however, for this study we focus on the approximately 144,000 ESI level 2 through level 5 walk-in patients. We do not study ESI level 1 patients because these patients do not abandon. However, we do include ESI level 1 patients in all relevant census measures in the analysis.

Available patient-level information includes demographics, clinical information, and time stamps. Patient demographics include age, gender, and insurance classification (private, Medicare, Medicaid, or none). Clinical information includes the patient's pain level on a 1 to 10 scale (10 being most severe), the chief complaint as recorded by the triage nurse, and a binary variable indicating whether the patient had any diagnostic tests, such as labs or X-rays, ordered at triage. Time stamps include time of arrival, time of being called for treatment, 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.

Empirical analysis of queue 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, also known as the offered wait time (Mandelbaum and Zeltyn 2013). Only the minimum of these two (actual wait time or actual abandonment time) is usually ever realized, leading to censored data. However, in the study hospital, the situation is different. For patients who abandon, we do not observe when they abandon, but for most of these patients, we do observe when the nurse first attempts to find the patient in the waiting room to bring him or her back to a treatment room. We refer to this as the first-call time, *CALL*. So although we do not know when the patient left, we do know how long his or her wait would have been had he or she stayed for service, the offered wait.

To formally define the offered wait, *OWAIT*, we consider three subgroups of patients: those who remain for treatment, those who abandon and we observe the first-call time, and those who abandon and the first-call time is missing.

Table 1 Summary Statistics

	ESI level 2	ESI level 3	ESI level 4	ESI level 5
Age (years)	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)
Triage diagnostics (%)	28 (0.003)	29 (0.002)	22 (0.002)	4 (0.002)
FastTrack (%)	2 (0.001)	3 (0.001)	68 (0.002)	67 (0.005)
Wait time (hours)	1.0 (0.01)	1.9 (0.01)	1.3 (0.01)	1.3 (0.01)
Service time (hours)	3.7 (0.02)	4.0 (0.01)	1.8 (0.01)	1.2 (0.01)
Census upon arrival	12.3 (0.05)	10.5 (0.03)	10.7 (0.04)	10.3 (0.07)
LWBS (%)	1.5 (0.001)	9.0 (0.001)	4.3 (0.001)	6.8 (0.002)
N	27,538	65,773	39,878	10,509

Note. Means are shown, with the standard errors of means in parentheses.

For patients who remain, their offered wait is just their actual wait, which we calculate directly from the time stamps. For most patients who abandon, we observe that the first-call time and the offered wait can likewise be calculated directly from the time stamps. Thus the general expression of offered wait is

$$OWAIT_i = \begin{cases} ROOM_i - ARRIVE_i \\ \quad \text{if patient stays,} \\ CALL_i - ARRIVE_i \\ \quad \text{if patient abandons,} \end{cases} \tag{2}$$

where the variable *ROOM* is the time stamp of when a patient was taken out of the waiting room and put in a treatment room, and the variable *ARRIVE* is the time stamp of when a patient first arrived to the ED.

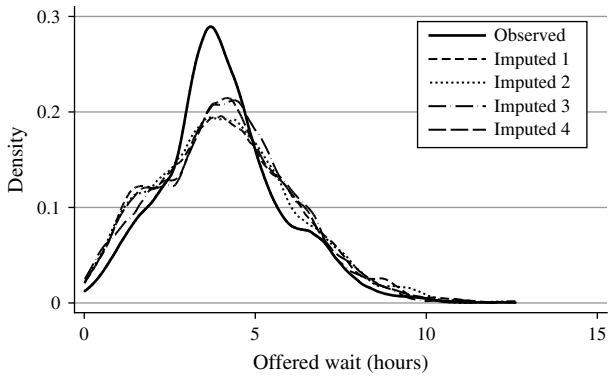
The data contain approximately 2,000 observations of patients who abandon but for whom we do not observe a first-call time leading to missing offered wait values (approximately 1.5% of the population and 20% of all patients who abandon). These missing data occur when the nurse fails to record the call time in the electronic patient-tracking system. In this paper we use the multiple imputation method to impute the missing values because it does the best job of matching the distribution of the missing data. We describe and compare other imputation methods in §8.1. All methods give similar results.

First developed by Rubin (1977, 1980), and with similarities to bootstrapping procedures, multiple imputation uses a first-stage model to predict the distribution of the missing values based on the observed values of all other variables (including the dependent variable) and uses Monte Carlo simulation to generate multiple random samples of the imputed values. The desired econometric analysis (probit, in our case) is then performed on each generated data set and the results combined across the data sets to determine coefficient point estimates and standard errors that are adjusted for the uncertainty as a result of the missing data (Rubin 1996, Cameron and Trivedi 2005, Section 27.7).

An important feature of multiple imputation is that its effectiveness is not based on the accuracy of the imputed values themselves, but rather on the accuracy of the distribution of the imputed values. Although it is impossible to assess the distribution accuracy directly (the distribution of the missing values is unknown), we compare the distribution of observed offered waits (patients with first-call times) to the distribution of imputed offered waits (patients without first-call times) for those who abandon. Figure 2 shows the kernel density plots of observed offered waits and four arbitrary imputed samples. We observe that the distributions match fairly well; *t*-tests (not shown) indicate that the sample means are not significantly different, whereas the sample standard deviations of the imputed samples are slightly larger than the observed sample standard deviation (e.g., 2.1 versus 1.8 hours). Another factor that impacts the effectiveness of multiple imputation is the proportion of missing data, which in the present case is quite low. The similarity of distributions and the low proportion of missing data together bolster the case for using multiple imputation.

To calculate the waiting room census measure, we divide the study period into a series of 15-minute intervals labeled t, and we use the patient

Figure 2 Kernel Density Plots of Observed and Imputed Wait Times for ESI Level 3 Patients Who Abandon



visit time stamps 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 ($INTERVAL\_CENSUS_{t,T}$, $T \in \{1, 2, 3, 4, 5\}$).

One impediment to estimating $INTERVAL\_CENSUS_t$ is that we must make an assumption about when abandoning patients abandon since we do not observe the abandonment time. In §8.2 we examine four abandonment timing assumptions: instant abandonment, abandonment after one hour, abandonment at the time the patient record is closed out, and abandonment at a random time between arrival and record closeout. All assumptions except instant abandonment yield similar results. Prior research shows that the true abandonment behavior is somewhere between the two extremes of the instant and closeout assumptions with a mean time to abandonment of one hour (Fernandes et al. 1994, Arendt et al. 2003). Therefore, in this paper, we assume abandonments occur deterministically one hour after arrival.

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 of ED treatment beds in use at the time of arrival, which is the number of patients in the treatment phase of the visit.

To test Hypothesis 3, we would ideally decompose $CENSUS_i$ into those patients whom patient i perceives to be more or 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 level 3 patient, patients in the waiting room of ESI levels 1–3 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, 2B, and 2C and 4A, 4B, and 4C are constructed based on the patient time stamps. For each patient visit, we calculate the number of arrivals ($ARRIVALS_i$) and departures ($DEPART_i$) that occur within the first hour after patient i's arrival. Furthermore, 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 ($ARRIVALS_{i,T}$, $DEPART_{i,T}$, $NONJUMP_{i,T}$, $JUMP_{i,T}$, $T \in \{1, 2, 3, 4, 5\}$).

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:

- $ARRIVALS\_AHEAD_i$: arriving patients with a higher priority than patient i
- $ARRIVALS\_BEHIND_i$: arriving patients with an equal or lower priority than patient i
- $DEPART\_AHEAD_i$: departing patients with an equal or higher priority than patient i
- $DEPART\_BEHIND_i$: departing patients with a lower priority than patient i
- $NONJUMP\_AHEAD_i$: departing patients with an equal or higher priority than patient i and who arrived before patient i
- $NONJUMP\_BEHIND_i$: departing patients with a lower priority than patient i and who arrived before patient i
- $JUMP\_AHEAD_i$: departing patients with a higher priority than patient i and who arrived after patient i
- $JUMP\_BEHIND_i$: departing patients with an equal or lower priority than patient i and who arrived after patient i

Note that the "jump"/"nonjump" language indicates relative arrival timing only, whereas the "ahead"/"behind" language indicates relative position in the priority queue, which is a function of both arrival timing and priority level. At first thought, it may appear as though jump-behind departures should not occur. However, they do indeed occur for several reasons. As noted in §2, FCFS is sometimes violated by the charge nurse for clinical or load-leveling reasons. The other major cause of jump-behind departures is the FastTrack. For example, when the FastTrack is open, an ESI level 3 patient may wait a long time to be seen in the regular ED while several less acute ESI level 4 and level 5 patients arrive and are served in the FastTrack. These arrivals and the subsequent departures are recorded as behind-arrivals and jump-behind departures.

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 ($ARRIVALS_i$, $DEPART_i$, $NONJUMP_i$, $JUMP_i$, 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 one 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 departures. 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. We focus first on the binary regression model used for the primary analysis and then on the multiple imputation framework in which it is embedded.

6.1. Binary Regression Model

The observed outcome variable is whether or not a patient abandons the waiting room. This calls for the use of a model of binary choice such as the logit, probit, skewed logit, or complimentary log log (Greene 2012, Section 17.2; Nagler 1994). Such models fit naturally with the utility-theory choice framework presented in §4 because they can be interpreted as modeling the difference in utility between two possible actions (stay or abandon) 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. In this setting, the estimated coefficients can be interpreted as the impact of a given independent variable on the utility function presumably through a change in the patient's estimated residual wait time. Since ε is stochastic, these models can only predict a probability of choosing one action over the other.

Selecting the best binary-choice model a priori is difficult because each has theoretical or practical advantages and disadvantages. We test and compare the several binary models (not shown) and find that for the coefficients of interest, all models provide similar results. In this paper, we present the results from the probit model.

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

$$\begin{aligned}\text{Prob}(LWBS_i = 1 \mid \mathbf{x}) &= \Phi(\beta_0 + \beta_1 OWAIT_i + \beta_2 CENSUS_i \\ &\quad + \beta_3 OWAIT_i \times CENSUS_i + \mathbf{X}_i\boldsymbol{\beta}_P + \mathbf{Z}_i\boldsymbol{\beta}_T), \quad (3)\end{aligned}$$

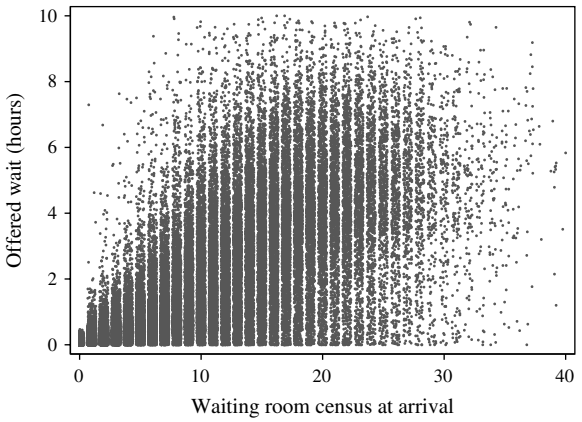
where $\Phi(\cdot)$ represents the standard normal cumulative distribution function; $\mathbf{X}_i$ is a vector of patient-visit specific covariates including age, gender, insurance type, chief complaint, pain level, and a dummy variable indicating if a patient received any diagnostic testing orders at triage; and $\mathbf{Z}_i$ 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. The subscripts P and T on the two beta vectors indicate that $\boldsymbol{\beta}_P$ is a vector of coefficients related to patient characteristics and $\boldsymbol{\beta}_T$ is a vector of coefficients related to time 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 ESI level 2 and level 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, because of the nonlinear nature of the probit model, the interaction coefficient cannot be interpreted in such a straightforward way. We discuss this interpretation further in §7.2.1.

The *OWAIT* variable is a bit different from all the other queue status variables in the model in that it is not actually observed by the patient. Even for patients who enter the service phase, 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 amount of time a patient can spend in the system deciding whether to stay or abandon. The Erlang-A model incorporates this idea that the longer a person is in the system, the higher her total probability of abandoning. Thus, the *OWAIT* variable is a control variable that picks up this effect that patients who are given the opportunity to be in the system longer are more likely to abandon, even though the actual offered wait value is not observed by the patient.

Our identification strategy is based on the assumption that *OWAIT* and *CENSUS* are not perfectly correlated and both contain exogenous variation. 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). 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 3, which shows the scatterplot of *OWAIT* and *CENSUS* (waiting room census at arrival) for ESI level 3 patients. Note that

Figure 3 Scatterplot of Offered Wait and Census for ESI Level 3 Patients



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

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 (VIFs) 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. Adding the *OWAIT* × *CENSUS* interaction term leads to a mild increase in the VIF of the *CENSUS* variable, but it remains well below 10. Furthermore, a comparison of the standard errors of estimating Equation (3) with and without the interaction term (not shown) shows that the standard errors remain quite small and stable, indicating that multicollinearity is not a concern.[2]

Finally, 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, Section 17.3.1). 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.

[2] We thank an anonymous reviewer for raising the issue of the interaction term potentially creating a collinearity issue and for pointing out that although mean-centering can lead to reduced VIFs, it does not reduce the standard error of the point estimates. We refer the reader to Echambadi and Hess (2007).

6.2. Multiple Imputation Framework

As mentioned in §5, multiple imputation is a multistep process that first generates multiple imputed samples for the missing data, analyzes the completed data sets, and then combines the results. We refer the reader to Rubin (1996) and (Cameron and Trivedi 2005, Chap. 27) for a detailed discussion of multiple imputation and the underlying theory. The offered wait variable that contains missing data is highly skewed and bounded below at zero, and thus we use predictive mean matching to generate the imputed values rather than the more typical linear regression method (Little 1988, Schenker and Taylor 1996). The imputation process has five steps:

Step 1. Fit a linear regression model of offered wait on all the other variables in Equation (3), including the dependent variable *LWBS*, using only observations that are complete. This provides estimated coefficients and variance of the prediction model.

Step 2. Generate a Monte Carlo sample of the prediction model coefficients and variance based on the joint posterior distribution informed by Step 1 (starting with a uninformative prior such as the Jeffreys prior).

Step 3. Using these new coefficients, calculate the predicted values for all observations, missing and nonmissing.

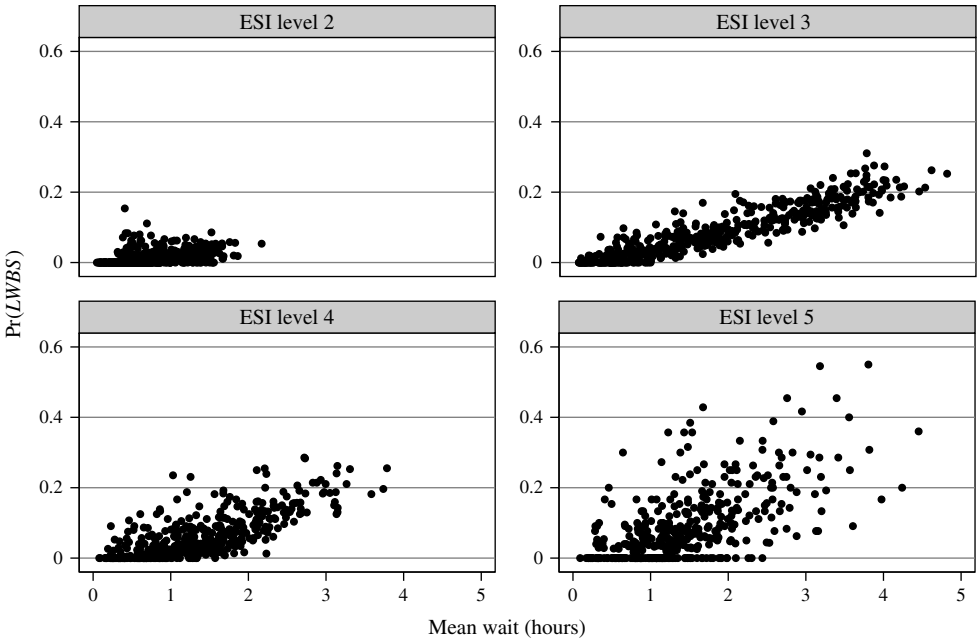
Step 4. For each incomplete observation, find the k complete observations with the closest predicted values and assign the actual value from one of those k observations to the incomplete observation, choosing at random.

Step 5. Repeat Steps 2–4 to generate M complete imputed data sets.

We set k to the typical value of 3 (Schenker and Taylor 1996). We generate 10 imputed data sets for each analysis ($M = 10$), which is more than sufficient given the low proportion of missing data. Other values of k and M produce similar results.

Once the M imputed data sets are generated, the primary probit analysis is done as described in §6.1. Analysis is repeated on each of the imputed data sets separately and the results are combined. The combined coefficient point estimate is calculated as the mean of the point estimates from each of the M data sets. The combined variance of each coefficient is calculated as the mean of the variance from each imputed sample plus the variance of the coefficient point estimates across the imputed samples. Thus the variance and related standard errors capture both within-imputation variability and between imputation variability (Rubin 1996).

Figure 4 **Pr(*LWBS*) vs. Wait Time**



Notes. Each point is year/day/hour. Points that represent fewer than 10 observations are excluded.

7. Results

7.1. Overview Graphs

Following the example of Zohar et al. (2002), we begin by using scatterplots to visualize the relationship between abandonment and wait time.[3] Zohar et al. prove via Little's law that a $G/G/m{+}M$ queue with exponential patience distribution exhibits an increasing linear relationship between the expected wait time and the probability of abandonment. Brandt and Brandt (2002) extend this result to Erlang-A-type queues with state-dependent arrival and service rates and general patience distributions.

Figure 4 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 who arrived on Tuesdays in 2009 during the 4 P.M. hour. Each graph has approximately 504 points (3 years × 7 days × 24 hours = 504). However, points that represent fewer than 10 observations have been dropped. For example, there are not many ESI level 5 patients at 4 A.M on Mondays in 2009, so that point has been dropped. Each subplot of Figure 4 displays a single triage or ESI level. In summary, each dot shows for patients who arrived at a given year/day/hour their average wait time and the percentage of people who abandoned.

[3] Similar analysis appears in Lucas et al. (2014) in support of setting target wait times in EDs.

We observe several interesting features in Figure 4. First, there is a linear increasing trend for all triage levels, as predicted by Zohar et al. (2002) and Brandt and Brandt (2002) (see Table 2 for the slope of a linear best-fit line). This relationship to wait time can be thought of as the baseline abandonment level for a given offered wait that is controlled for by the inclusion of offered wait in the regression models of Equation (3).

Second, the slope of the linear fit decreases with acuteness (see Table 2). This suggests that sicker patients are less influenced by wait time. In terms of the utility model, this behavior is consistent with sicker patients having a higher expected reward for service, and thus a larger change in expected residual wait time is necessary to cause a sick patient to abandon.

The third feature we observe in Figure 4 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 4. Additionally, from the R^2 values in Table 2, we conclude that the mean wait

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

ESI level	Slope	RMSE	R^2
2	0.021 (0.002)	0.016	0.238
3	0.057 (0.001)	0.026	0.875
4	0.065 (0.003)	0.033	0.598
5	0.079 (0.005)	0.071	0.370

Note. Standard errors are shown in parentheses.

time is a very good predictor of abandonment probability for ESI level 3 patients. For ESI level 4 and level 5 patients, however, there appear to be other factors driving abandonment beyond just wait time. The results for ESI level 2 appear somewhat different. Although ESI level 2 displays a positive linear trend with little dispersion (significant positive slope and low RMSE), the model has the lowest R^2 value, further indicating that wait time explains very little of the variation in ESI level 2 patient 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 §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 §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, which we cannot do as easily with the consolidated data in §7.1. For clarity, we focus on results for ESI level 3 in §§7.2.1 and 7.2.2. We select ESI level 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 §7.2.3.

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 because they represent a change in the linear z-score predictor as a result of a change in an independent variable. The first-order terms of *OWAIT* 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.

Table 3 Effect of Observed Variables on Pr(*LWBS*) (ESI Level 3)

	Model 1	Model 2	Model 3
OWAIT	0.25*** (0.01)	0.19*** (0.01)	0.18*** (0.01)
CENSUS	0.11*** (0.00)	0.10*** (0.00)	0.10*** (0.00)
OWAIT × *CENSUS*	−0.02*** (0.00)	−0.02*** (0.00)	−0.02*** (0.00)
ARRIVALS			0.01** (0.00)
NONJUMP DEPARTURES			−0.03*** (0.00)
JUMP DEPARTURES			−0.01 (0.01)
N	65,618	35,865	35,865

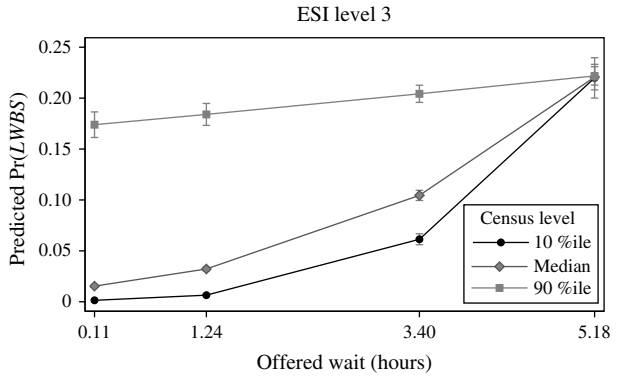
Notes. Cluster robust standard errors are in parentheses. Controls not shown include age, gender, insurance, pain, chief complaint, triage test, year, weekend, block of day, and the interaction term of weekend and block of day.
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

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 the value of all the other covariates. To get a sense of the magnitude of effects, we calculate the mean marginal effect (across patients) of both the offered wait and census variables at their respective median values of 1.24 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 0.6-percentage-point increase in abandonment for ESI level 3 patients. (We can alternatively describe the marginal impact of an additional person in the waiting room as being equivalent to a 18-minute increase in offered wait.) This supports Hypothesis 1 and shows that waiting patients are influenced by the waiting room census. A possible explanation is that as patients wait, they form an expected residual wait by multiplying an estimated interdeparture time by the number of people ahead of them in line. If the estimated interdeparture time changes as a result of observation of the system, the resulting change in the residual wait will be magnified when the census is large.

Lu et al. (2013) estimate 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, similar to our estimated marginal effect of 0.6 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 5 shows the predicted abandonment probabilities at various levels of offered wait time and census. Offered wait time is on the x axis, and the four test points (0.11, 1.24, 3.40, and 5.18 hours) are the 10th, 50th, 75th, and 90th percentiles for ESI level 3 patients, respectively. Each line on the graph represents the predicted probability of abandonment for a given census level. The three lines indicate the 10th, 50th, and 90th percentile census levels (1 person, 10 people, 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

Figure 5 **Predicted Pr(*LWBS*) as a Function of Offered Wait and Census**



Notes. Error bars indicate the 95% confidence interval of the estimated point. Some bars are so narrow as to not be visible in the figure. %ile, percentile.

patients are responding to the census level as well as the wait time. For example, a patient who arrives when the waiting room is relatively empty and experiences a 1.24-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 5 shows that patients respond to both increasing offered wait time and waiting room census with increased abandonment.

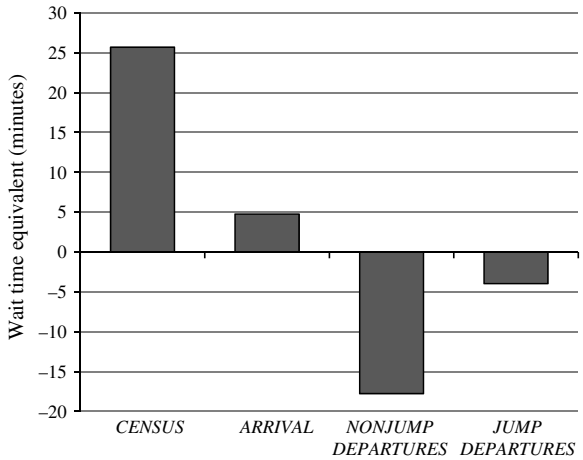
The large gap between the median and 90th percentile census levels even for very short waits suggests that large crowds lead to rapid abandonment even when the actual wait time is low. This also explains why the slope of the 90th percentile census line is relatively flatter. People are likely abandoning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

To examine Hypotheses 2A, 2B, and 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 nonjump departures is significant and negative. This supports Hypothesis 2B, that observing nonjump departures leads to reduced abandonment, presumably because waiting patients view these departures as a good sign of processing speed and progress toward service.

Model 3 of Table 3 adds in variables for the number of arrivals to the ED and the number of departures into service (split into nonjump and jump departures). The positive and significant coefficient on arrivals supports Hypothesis 2A, that arrivals lead to more abandonments. The coefficient on nonjump departures is significant and negative. This supports Hypothesis 2B, that observing nonjump departures leads to reduced abandonment, presumably because waiting patients view these departures as a good sign of processing speed and progress toward service. The insignificant effect of jump departures suggests that any positive information about system speed is largely negated by the fact that the patient is getting jumped and is not moving closer to the head of the line. However, testing Hypothesis 2C requires testing the difference between the coefficients *NONJUMP DEPARTURES* and *JUMP DEPARTURES*. A one-sided t-test comparing the nonjump and jump coefficients rejects the null hypothesis of the *NONJUMP DEPARTURES* coefficient being greater than or equal to the *JUMP DEPARTURES* coefficient ($p < 0.01$), therefore supporting Hypothesis 2C, that the *JUMP DEPARTURES* coefficient is significantly larger (less negative). In terms of marginal effects, observing an arrival increases abandonment by 0.2 percentage points, and observing a nonjump departure reduces abandonment by 0.7 percentage points. Figure 6 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 five minutes of offered wait time. Similarly, observing a nonjump departure has the same impact on abandonment as a 18-minute reduction in offered wait time.

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



Note. The *JUMP DEPARTURES* marginal effect estimate is statistically insignificant.

Table 4 Effect of Inferred Variables on Pr(*LWBS*) (Probit, ESI Level 3)

	Model 1		Model 2	
OWAIT	0.25***	(0.01)	0.18***	(0.01)
CENSUS_AHEAD	0.14***	(0.00)	0.13***	(0.00)
CENSUS_BEHIND	0.03***	(0.01)	0.03***	(0.01)
OWAIT × *CENSUS_AHEAD*	−0.03***	(0.00)	−0.02***	(0.00)
OWAIT × *CENSUS_BEHIND*	−0.01***	(0.00)	−0.01***	(0.00)
ARRIVALS_AHEAD			0.05***	(0.01)
ARRIVALS_BEHIND			−0.00	(0.00)
NONJUMP_AHEAD			−0.03***	(0.00)
NONJUMP_BEHIND			−0.01*	(0.01)
JUMP_AHEAD			−0.06***	(0.02)
JUMP_BEHIND			0.01	(0.01)
Marginal effects				
CENSUS_AHEAD	0.012***	(0.001)	0.013***	(0.001)
CENSUS_BEHIND	0.001	(0.001)	0.002**	(0.001)
N	65,618		35,865	

Notes. Clustered robust standard errors are in parentheses. Controls not shown include age, gender, insurance, pain, chief complaint, triage test, year, weekend, block of day, and the interaction term of weekend and block of day.
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

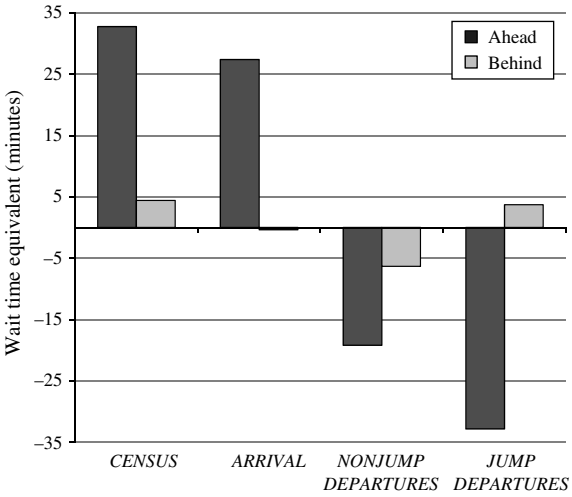
In summary, patients respond to what they observe, and the magnitudes of their responses are similar in magnitude to 5 to 25 minutes of change in waiting time.

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 other patients who are ahead of or behind them 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 of Table 4 is analogous to Model 1 of Table 3 but with the census variable split into ahead and behind components as described in §5. It is estimated on the full sample. Although the coefficients of *CENSUS_AHEAD* and *CENSUS_BEHIND* are both statistically significant, only the marginal effect of *CENSUS_AHEAD* is significant, whereas the marginal effect of *CENSUS_BEHIND* is not. This supports Hypothesis 3, that patients respond more strongly to those ahead in the priority queue than those behind. In Model 2 of Table 4, which is analogous to Model 3 of Table 3 but with the census and flow variables split into their respective ahead and behind components, the marginal effects of both *CENSUS_AHEAD* and *CENSUS_BEHIND* are positive and significant, but a one-sided test of the marginal effects shows that the *CENSUS_AHEAD* effect is significantly larger than the *CENSUS_BEHIND* effect ($p < 0.01$), further supporting Hypothesis 3.

Looking at the flow variables in Model 2, we compare the coefficients of each ahead/behind pair and find that the values are significantly different and that the ahead component always has a larger magnitude than the behind component (p-values < 0.01 for all three pairs of coefficients; two-sided tests of coefficient equality are also all rejected with p-values < 0.01). This supports Hypotheses 4A, 4B, and 4C.

Figure 7 Magnitude of Marginal Effects in Equivalent Minutes of Offered Wait Time



Notes. All "ahead" marginal effects are significant at the 5% level. *CENSUS_BEHIND* is the only significant "behind" marginal effect.

Similar to Figure 6, Figure 7 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 larger than for the observed variables in Figure 6. We note that although the point estimates of the *NONJUMP_AHEAD* and *JUMP_AHEAD* coefficients seem quite disparate (−20 minutes and −33 minutes, respectively), 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 who interfere with their experience. Although the directions of the effects are all as expected, this result is noteworthy because it shows that patients are indeed inferring relative priority information by observing the other patients.

7.2.3. Results Across Triage Levels. Table 5 shows the results of the most detailed 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. The main difference between the models is that

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

	Model 1: ESI level 2	Model 2: ESI level 3	Model 3: ESI level 4	Model 4: ESI level 5
OWAIT	0.34*** (0.04)	0.18*** (0.01)	0.28*** (0.02)	0.30*** (0.03)
CENSUS_AHEAD	0.12*** (0.02)	0.13*** (0.00)	0.04*** (0.01)	0.04*** (0.01)
CENSUS_BEHIND	0.00 (0.01)	0.03*** (0.01)	0.03 (0.03)	
OWAIT × *CENSUS_AHEAD*	−0.03*** (0.01)	−0.02*** (0.00)	−0.01*** (0.00)	−0.01*** (0.00)
OWAIT × *CENSUS_BEHIND*	0.00 (0.00)	−0.01*** (0.00)	−0.00 (0.01)	
ARRIVAL_AHEAD	0.05 (0.17)	0.05*** (0.01)	0.02*** (0.01)	0.03** (0.01)
ARRIVAL_BEHIND	0.01 (0.01)	−0.00 (0.00)	0.00 (0.01)	−0.04 (0.03)
NONJUMP_AHEAD	−0.04 (0.02)	−0.03*** (0.00)	−0.02*** (0.01)	−0.03** (0.01)
NONJUMP_BEHIND	−0.01 (0.01)	−0.01* (0.01)	−0.04 (0.03)	
JUMP_AHEAD	0.06 (0.21)	−0.06*** (0.02)	0.00 (0.02)	0.01 (0.03)
JUMP_BEHIND	0.01 (0.03)	0.01 (0.01)	0.08 (0.05)	0.36** (0.16)
Marginal effects				
CENSUS_AHEAD	0.004***(0.001)	0.013***(0.001)	0.002***(0.000)	0.004***(0.001)
CENSUS_BEHIND	0.000 (0.000)	0.002** (0.001)	0.002 (0.002)	
N	9,118	35,865	19,775	5,259

Notes. Clustered robust standard errors are in parentheses. Controls not shown include age, gender, insurance, pain, chief complaint, triage test, year, weekend, block of day, and the interaction term of weekend and block of day.

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

Models 1, 3, and 4 have fewer significant coefficients, which is not surprising given the smaller sample sizes for these models compared with Model 2 (ESI level 3).

ESI level 5 is the most dissimilar of the four models. First, the variables *CENSUS_BEHIND* and *NONJUMP_BEHIND* are not included in the ESI level 5 model because level 5 is the lowest priority level. Second, the coefficient for jump-behind departures is positive and significant. This means that ESI level 5 patients are more likely to abandon when they observe patients of the same acuteness arrive after them and be served before them. The marginal effect is approximately a five-percentage-point increase in probability of abandonment for each observed jump-behind departure. This effect is likely large with this group of patients because their utility reward for service is so low that the perceived unfairness of a jump-behind departure far outweighs the positive speed of service signal.

8. Robustness Tests

8.1. Alternatives for Handling Missing Data

As mentioned in §5, the first-call time stamp (*CALL*) is missing for about 2,000 patients who abandon the ED. Because the missing data condition only exists for patients who abandon, we cannot simply drop the observations with missing data. In this paper, we use the multiple imputation method of Little (1988) and Rubin (1996) to impute the missing data as described in §6.2.

A potential weakness of the multiple imputation method is that it is based purely on the ability of the observed data to predict the missing data. In this sense, it is rather "blind" in that it does not take advantage of any knowledge of the actual data generation mechanism being studied. In the study at hand, the data-generating mechanism is a multiclass queue, and we can use the logic of how the queue behaves, rather than statistics, to generate imputed values for the missing data. We develop four offered wait imputation methods based on the operational behavior of the queue, as follows:

MeanWait $i \pm 1$: $OWAIT_i = \frac{1}{2}(WAIT_{i-1} + WAIT_{i+1})$ (patients $i-1$ and $i+1$ must be of the same triage class as patient i and not abandon)

TimeStamp $i+1$: $OWAIT_i = READY_{i+1} - ARRIVE_i$ (patient $i+1$ must be of the same triage class as patient i)

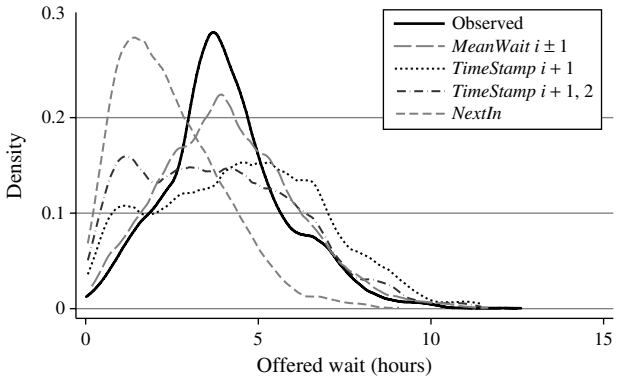
TimeStamp $[i+1, i+2]$: $OWAIT_i = \min[READY_{i+1}, READY_{i+2}] - ARRIVE_i$ (patients $i+1$ and $i+2$ must be of the same triage class as patient i)

NextIn: $OWAIT_i = \min[READY_{i+1}, READY_{i+2}, \ldots, READY_N] - ARRIVE_i$ (patients $i+1, \ldots, N$ must be of the same triage class as patient i)

Let $READY_i = \min[CALL_i, ROOM_i]$. The *MeanWait* $i \pm 1$ method calculates the mean of the actual wait times of the chronologically adjacent patients who remained for service. The method is predicated on the fact that there is serial correlation across wait times. However, this method is negatively biased because it only uses information from patients who stayed for treatment. Since wait time is a driver of abandonment, on average, the patients who remained have lower offered wait times than those who abandoned.

The other three estimating methods are all based on the assumption that, in general, if patient i of triage class T is not present when called, the nurse will move on to subsequent patients in the queue of

Figure 8 **Kernel Density Plot of Queue-Based Offered Wait Imputation Methods for ESI Level 3 Patients**



the same triage class. If this "next-in-line" rule were always followed, then the *TimeStamp* $i+1$ method would be sufficient, and the *READY* time stamp of patient $i+1$ should be imputed to patient i. However, in practice, the next-in-line rule is not always followed, and thus the remaining two methods attempt to correct for this by taking the minimum time stamp of the subsequent two patients or all remaining patients, respectively. If it takes a few moments for the nurse to switch from calling patient i to a subsequent patient, then these measures could have a positive bias.

Figure 8 shows the kernel density plots of these offered wait estimates compared with the observed offered wait times for ESI level 3 patients who abandon, similar to Figure 2. The *MeanWait* $i \pm 1$ method most closely matches the observed distribution, whereas none of the other methods appears to match very closely.

Table 6 shows the descriptive statistics of the deviations of the estimated offered wait value from the actual value offered wait value for the approximately 6,900 patients who abandoned but for whom we have a first-call time. The *MeanWait* $i \pm 1$ method has the smallest standard deviation of differences but has a negative bias, as predicted. The *TimeStamp* $[i+1, i+2]$ method is the only method that is unbiased (the point estimate of 0.04 is statistically insignificant). All four methods show fairly wide distributions of deviations from the actual value, which suggests that these queue mechanics-based methods introduce a lot of measurement noise into the model.

Table 6 **Comparison of Deviation of Estimated Offered Wait from the Actual Wait for Patients Who Abandon (Estimate−Actual)**

	MeanWait $i \pm 1$	*TimeStamp* $i+1$	*TimeStamp* $[i+1, i+2]$	*NextIn*
Mean	−0.22	0.55	0.04	−1.03
SE of mean	0.02	0.03	0.03	0.03
SD	1.96	2.51	2.59	2.46
10th percentile	−2.51	−2.47	−3.20	−4.02
Median	−0.21	0.33	0.07	−0.83
90th percentile	2.12	3.37	2.88	1.45

Note. All values are shown in hours.

Table 7 **Waiting Room Census: Comparison of Census Estimation Assumptions**

	Instant	One-hour	Closeout	Random
Mean	8.12	8.46	9.62	8.86
SE of mean	(0.02)	(0.02)	(0.03)	(0.02)
SD	7.05	7.42	8.89	7.97
10th percentile	0	0	0	0
Median	7	7	7	7
90th percentile	18	19	23	21

Note. All values are shown in number of patients.

We repeat all the analyses in this paper using each of these four alternative offered wait estimation methods (not shown). Perhaps surprisingly, none of the results changed appreciably. This is likely because the missing data affect such a small portion of the data—only about 1.5%.

8.2. Census and the Abandonment Timing Assumption

The waiting room census ($INTERVAL\_CENSUS_t$) is calculated from the arrival and departure time stamps of each patient visit. However, for patients who abandon the queue, we do not observe when they leave the waiting room. Therefore, we must make an assumption about when to remove them from the waiting room census. We test the following four abandonment timing assumptions:

- Instant: LWBS patients abandon instantly and are never included in the waiting room census.
- One-hour: LWBS patients abandon one hour after arrival.
- Closeout: LWBS patients abandon at the final time stamp of the observation, which closes out that patient visit.
- Random: LWBS patients abandon at a point randomly chosen with continuous uniform probability between the arrival time and the record closeout time.

Table 7 provides descriptive statistics of the census measure under each of the above assumptions. By construction, instant abandonment leads to the lowest estimated census, and abandonment at closeout leads to the highest census estimate, with a mean difference between these two measures of about 1.5 patients. This difference is largest during periods of high patient abandonment (e.g., weekday evenings when the ED is crowded). During these periods, assuming actual abandonment behavior is somewhere between the two behavioral extremes, the instant abandonment assumption underestimates the census while the closeout assumption overestimates the census.

We replicate the study analyses under each of the four abandonment assumptions (not shown) and find that the results are quantitatively similar and qualitatively identical for the one-hour, closeout, and random assumptions. However, under the instant abandonment assumption, the independent variable of jump departures becomes negative and significant. This occurs because the instant abandonment assumption systematically underestimates the census measure at times of high load (when most abandonments occur), and these conditions happen to be negatively correlated with jump departures. Thus the jump departures variable erroneously appears to have an impact on abandonment because of its mild correlation with the mismeasured census variable. Assuming abandonment occurs one hour after arrival, rather than instantaneously, can sufficiently alleviate this spurious result.

As mentioned in §5, prior research shows that abandonment behavior is somewhere between the two behavioral extremes of the instant and closeout assumptions with a median time to abandonment of one hour (Fernandes et al. 1994, Arendt et al. 2003), and thus the one-hour and random abandonment assumptions are likely better estimates of the true census. We use the one-hour abandonment assumption throughout the paper because of its simplicity.

To further check the robustness of our findings to measurement error in the census, we repeat the analyses on the subsample of patients for whom there is no LWBS-induced measurement error. That is, we only include patients who arrived when we know that there are no abandoning patients in the waiting room.[4] Although this is not at all a random subsample (this condition tends to occur in the morning and when offered wait and census is low) and the sample size is dramatically reduced (37,062 versus 65,618), we find that the estimated coefficients of interest remain significant and are quite similar in magnitude to the main results shown in the paper.

8.3. Potential for Shifts in Patient Mix and Strategic Patients

It is possible that patients differ on dimensions unobservable to the researcher (e.g., tolerance for waiting or crowds) and that these different types of patients tend to come to the ED at different times of the day or days of the week. Likewise, it is possible that some patients are strategic and plan their arrivals for times when they expect the wait to be low, and these strategic patients may have different abandonment characteristics than nonstrategic patients. These potential systematic shifts in patient mix would create an endogeneity problem that would bias the results. In the main model, we control for this possibility by including variables that control for time of day (by four hour blocks: 4 A.M–8 A.M, 8 A.M–12 P.M., etc.) and for whether a patient visit is on a weekday or weekend. These control variables are generally significant, showing that there are systematic shifts by time of day and day of week. We also test other time/day control parameterizations (day of week, hour or day, work shift, etc.) and find that the results are largely unchanged.

Controlling for time of day and day of week allows for shifts in abandonment probability but does not allow for the coefficients of interest (census, arrivals, etc.) to vary by time of day. Interacting the key variables with the time variables would allow for this but makes interpretation difficult because there are so many time dummies. Alternatively, we repeat the main analyses of the paper on subsamples of the data limited by time of day (not shown). For example, we run the analyses on patients who arrive between midnight and 8 A.M, the least busy hours of the day. Likewise, we run the analysis on patients who arrive between noon and 4 P.M., the busiest block of the day. The coefficient point estimates remain largely unchanged. However, because of the much smaller sample size, some coefficients are no longer significant. Thus it appears that the results of the study are not driven by unobserved shifts in patient mix over the course of the day.

9. Discussion and Conclusion

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 this impacts the probability of abandoning the queue. Ours is among the first studies to show customers responding to the actual functioning of the queue. We expand on prior work showing that the queue length (in our study, waiting room census) impacts behavior separate from wait time. This shows that in queues that are at least partially visible, Erlang-A-type models do 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 that new arrivals may be served first. Furthermore, waiting patients infer the relative priority status of

[4] We thank an anonymous referee for suggesting this robustness check.

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 who will generally be served first. All of these effects are consistent with a utility-framework view of patients updating their expected residual wait time in response to what they observe and experience *during* the waiting period. This is managerially relevant for any organization that wants to manage customer queue abandonment.

The main econometric model used in this paper is the probit model as a result of the binary nature of the key outcome variable: whether a patient abandons the queue. This model is particularly useful for this work because it has a natural utility framework interpretation that aligns with the patient-choice utility model that is the theoretical foundation of this work. The probit model allows us to estimate the marginal impact of observed events on the probability of abandonment while controlling for many other factors such as patient demographics and seasonality effects. One drawback of the probit model is that it is a static model, in that we treat each patient encounter as a single observation in the data set. In contrast, an alternative way to study queue abandonment would be with a dynamic model such as a duration model with time-varying covariates (Wooldridge 2010, Section 22.2.3). This type of analysis would allow us to estimate the impact of an observed event (e.g., an arrival) on the hazard rate of abandonment. In short, duration analysis would allow us to study how the waiting experience impacts not just the *probability* of abandonment but also the *timing* of abandonment. Unfortunately, this is not possible with the available data because abandonment times are not observed. Although abandonment timing data are commonly available in call-center settings (Zohar et al. 2002, Mandelbaum and Zeltyn 2013), we are not aware of any research on visible queues with observed abandonment times. This is an excellent opportunity for future research to verify and extend our results if a setting can be found where abandonment times are captured. Possible settings include an ED that uses radio frequency identification tags to track waiting patients, retail stores that track customers' cell phones (Clifford and Hardy 2013), or shopping carts that track a customer's movements (Lindstrom 2011).

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 cannot accurately assess their need for treatment, then the hospital may want to withhold information. Finally, 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 recently began requiring hospitals to report ED performance measures such as median wait time, median length of stay, and LWBS percentage (Centers for Medicare and 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. The study hospital's 6.5% LWBS rate is much higher than the national mean level of 1.7% (Pham et al. 2009), and thus minimizing LWBS is an important issue for the study hospital.

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.

There has been a great deal of analytical work on queues with 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 improves either 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 invisible multiserver queues and find that providing more information can improve system performance with little customer

loss. Plambeck and Wang (2013) 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. Allon et al. (2011) consider the "what to tell customers" question under the assumption of strategic behavior by both customers and providers, and Jouini et al. (2011) explore 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.

Given this vast array of analytical work, more empirical work is needed to examine how people respond to various types of information during the waiting encounter. Motivated by the findings of the current work, future work should include a series of field 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. By contrast, it is unclear how providing more information will alter behavior. If the news appears to be bad (e.g., long wait time), abandonment may increase, but if the news appears to be good (e.g., short wait times despite long lines), abandonment may decrease. Field experiments will help determine whether these changes occur and what the net impact of the effects is. Lessons learned from such experiments will serve to improve both ED management and our general understanding of human queuing behavior.

Acknowledgments

The authors are grateful to Jesse Pines, George Washington University; Olan Soremekun, Jefferson University Hospital; and Christian Boedec, Hospital at the University of Pennsylvania, for their support and insight during this project. The authors thank the associate editor and three anonymous reviewers for their helpful comments that much improved the paper. Robert J. Batt's work was partially supported by grants from the Wharton Risk Management and Decisions Processes Center and the Fishman-Davidson Center for Service and Operations Management at the Wharton School.

References

Akşin Z, Ata B, Emadi S, Su CL (2013) Structural estimation of callers' delay sensitivity in call centers. *Management Sci.* 59(12): 2727–2746.

Allon G, Bassamboo A, Gurvich I (2011) "We will be right with you": Managing customer expectations with vague promises and cheap talk. *Oper. Res.* 59(6):1382–1394.

American College of Emergency Physicians (2012) Publishing wait times for emergency department care: An information paper. Report, American College of Emergency Physicians, Baltimore.

Arendt KW, Sadosty AT, Weaver AL, Brent CR, Boie ET (2003) The left-without-being-seen patients: what would keep them from leaving? *Ann. Emergency Medicine* 42(3):317–323.

Armony M, Mandelbaum A (2011) Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. *Oper. Res.* 59(1):50–65.

Armony M, Shimkin N, Whitt W (2009) The impact of delay announcements in many-server queues with abandonment. *Oper. Res.* 57(1):66–81.

Baccelli F, Hebuterne G (1981) On queues with impatient customers. Kylstra FJ, ed. *Performance '81: Proc. 8th Internat. Sympos. Computer Performance Modeling, Measurement, Evaluation* (North-Holland, Amsterdam), 159–179.

Batt RJ, Terwiesch C (2014) Doctors under load: An empirical study of state-dependent service times in emergency care. Working paper, University of Wisconsin–Madison, Madison.

Berry Jaeker J, Tucker AL (2012) Hurry up and wait: Differential impacts of congestion, bottleneck pressure, and predictability on patient length of stay. HBS Working Paper 13-052, Harvard Business School, Boston.

Bitran GR, Ferrer J-C, Rocha e Oliveira P (2008) Managing customer experiences: Perspectives on the temporal aspects of service encounters. *Manufacturing Service Oper. Management* 10(1): 61–83.

Bolandifar E, DeHoratius N, Olsen T, Wiler JL (2014) Modeling the behavior of patients who leave the emergency department without being seen by a physician. Chicago Booth Research Paper 12-14, University of Chicago, Chicago.

Brandt A, Brandt M (1999) On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation* 35(1–2):1–18.

Brandt A, Brandt M (2002) Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s+GI$ system. *Queueing Systems* 41(1/2):73–94.

Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* 100(469): 36–50.

Cameron CA, Trivedi PK (2005) *Microeconometrics: Methods and Applications* (Cambridge University Press, New York).

Centers for Medicare and 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 CW, Yom-Tov G, Escobar G (2014) When to use speedup: An examination of service systems with returns. *Oper. Res.* 62(2): 462–482.

Clifford S, Hardy Q (2013) Attention, shoppers: Store is tracking your cell. *New York Times* (July 15) A1–A3.

Echambadi R, Hess JD (2007) Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Sci.* 26(3):438–445.

Fernandes C, Daya MR, Barry S, Palmer N (1994) Emergency department patients who leave without seeing a physician: the Toronto hospital experience. *Ann. Emergency Medicine* 24(6): 1092–1096.

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3):208–227.

Gilboy N, Tanabe T, Travers D, Rosenau A (2011) *Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4, Implementation Handbook, 2012* (Agency for Healthcare Research and Quality, Rockville, MD).

Greene WH (2012) *Econometric Analysis*, 7th ed. (Prentice Hall, Upper Saddle River, NJ).

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

Hair JF Jr, Anderson RE, Tatham RL, Black WC (1995) *Multivariate Data Analysis*, 3rd ed. (Macmillan, New York).

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, International Series in Operations Research and Management Science, Vol. 59 (Springer, New York).

Haviv M, Ritov Y (2001) Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions. *Queueing Systems* 38(4):495–508.

Hobbs D, Kunzman SC, Tandberg D, Sklar D (2000) Hospital factors associated with emergency center patients leaving without being seen. *Amer. J. Emergency Medicine* 18(7):767–772.

Hsia RY, Asch SM, Weiss RE, Zingmond D, Liang LJ, Han W, McCreath H, Sun BC (2011) Hospital determinants of emergency department left without being seen rates. *Ann. Emergency Medicine* 58(1):24–32.

Huang T, Allon G, Bassamboo A (2013) Bounded rationality in service systems. *Manufacturing Service Oper. Management* 15(2): 263–279.

Hui MK, Tse DK (1996) What to tell consumers in waits of different lengths: An integrative model of service evaluation. *J. Marketing* 60(2):81–90.

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

Jouini O, Aksin Z, Dallery Y (2011) Call centers with delay information: Models and insights. *Manufacturing Service Oper. Management* 13(4):534–548.

Jouini O, Dallery Y, Aksin Z (2009) Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *Internat. J. Production Econom.* 120(2):389–399.

Kremer M, Debo LG (2013) Herding in a queue: A laboratory experiment. Chicago Booth Research Paper 12-28, University of Chicago, Chicago.

Larson RC (1987) Perspectives on queues: Social justice and the psychology of queueing. *Oper. Res.* 35(6):895–905.

Lindstrom M (2011) Shopping carts will track customers' every move. *Harvard Bus. Rev.* (blog) (December 9), http://blogs.hbr.org/2011/12/shopping-carts-will-track-cons.

Little RJA (1988) Missing-data adjustments in large surveys. *J. Bus. Econom. Statist.* 6(3):287–296.

Liu Y, Whitt W (2012) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* 60(6): 1551–1564.

Lu Y, Musalem A, Olivares M, Schilkrut A (2013) Measuring the effect of queues on customer purchases. *Management Sci.* 59(8): 1743–1763.

Lucas J, Batt RJ, Soremekun OA (2014) Setting wait times to achieve targeted left-without-being-seen rates. *Amer. J. Emergency Medicine* 32(4):342–345.

Mandelbaum A, Shimkin N (2000) A model for rational abandonments from invisible queues. *Queueing Systems* 36(1–3):141–173.

Mandelbaum A, Zeltyn S (2013) Data-stories about (im)patient customers in tele-queues. *Queueing Systems* 75(2–4):115–146.

Nagler J (1994) Scobit: An alternative estimator to logit and probit. *Amer. J. Political Sci.* 38(1):230–255.

Palm C (1943) *Intensitätsschwankungen im Fernsprechverkehr* (Esselte, Stockholm).

Pham JC, Ho GK, Hill PM, McCarthy ML, Pronovost PJ (2009) National study of patient, visit, and hospital characteristics associated with leaving an emergency department without being seen: Predicting LWBS. *Acad. Emergency Medicine* 16(10):949–955.

Plambeck EL, Wang Q (2013) Implications of hyperbolic discounting for optimal pricing and scheduling of unpleasant services that generate future benefits. *Management Sci.* 59(8): 1927–1946.

Polevoi SK, Quinn JV, Kramer NR (2005) Factors associated with patients who leave without being seen. *Acad. Emergency Medicine* 12(3):232–236.

Rubin DB (1977) Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Amer. Statist. Assoc.* 72(359):538–543.

Rubin DB (1980) Handling nonresponse in sample surveys by multiple imputations. Bureau of the Census Monograph, U.S. Department of Commerce, Washington, DC.

Rubin DB (1996) Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.* 91(434):473–489.

Schenker N, Taylor JMG (1996) Partially parametric techniques for multiple imputation. *Computational Statist. Data Anal.* 22(4): 425–446.

Shimkin N, Mandelbaum A (2004) Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Systems* 47(1–2):117–146.

Whitt W (1984) The amount of overtaking in a network of queues. *Networks* 14(3):411–426.

Wooldridge JM (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. (MIT Press, Cambridge, MA).

Zohar E, Mandelbaum A, Shimkin N (2002) Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Sci.* 48(4):566–583.