Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

Low-Acuity Patients Delay High-Acuity Patients in an Emergency Department

(Authors’ names blinded for peer review)

This paper provides empirical evidence that in an ED, arrival of an additional low-acuity patient significantly increases high-acuity patients’ wait times. To estimate that effect, we propose a method motivated by queueing theory and involving quasi-randomization in the wait time forecast at triage for low-acuity patients, which influences their propensity to leave without being seen. That small quasi-randomization (which is inexpensive to implement and poses little or no risk to patients) corrects a large bias in observational data. Empirically and using queueing theory, we identify mechanisms by which a low-acuity patient delays high-acuity patients: pre-triage delay and transition delay in preemption. Thus we identify ways to reduce high-acuity patients’ wait times: to limit preemption; to reduce the standard deviation or mean transition delay; to reduce unnecessary utilization of ED resources by low-acuity patients, e.g., by treating them outside beds or nudging ones that don’t need emergency treatment to choose to leave.

Key words: OM Practice, Service Operations, Empirical Research, Health Care Management, Queueing Models.
1. Introduction

Emergency Departments (EDs) operate according to the principle that high-acuity patients should “receive priority treatment, followed by the less urgent cases” (Iserson and Moskop 2007). Prioritizing high-acuity patients saves both lives and money, as even a few minutes of waiting to start treatment can increase mortality or recovery time and associated medical costs for high-acuity patients (Chan et al. 2008, Ferrer et al. 2014). To capture this operational principle, many Operations Management papers model EDs and supporting hospital facilities using preemptive priority queues with multiple acuity classes, as surveyed in (Fomundam and Herrmann 2007) and discussed in (Green 2006). In such models, a high-acuity patient’s wait time is not affected by the presence of low-acuity patients. Empirical emergency medicine literature supports that modeling assumption (Schull et al. 2007).

In contrast, this paper shows that low-acuity patients significantly increase wait times for high-acuity patients. Specifically, we observe via a causal analysis on data from four different hospitals’ EDs that the wait time for a high-acuity patient increases with the number of low-acuity patients who are in the system when he arrives. We then develop an innovative method to estimate the aggregate waiting externality imposed on all high-acuity patients by the arrival of one additional low-acuity patient, or “ExWAH” (“Extra Waiting for All High-acuity patients”), which is the appropriate metric to guide decisions and policies that influence the arrival rate of low-acuity patients. We find a significant ExWAH in observational data from the four different hospitals’ EDs, and validate that result through a quasi-randomized experiment at one hospital. Many papers attempt to detect causal effects from observational data, but such analyses can be subject to omitted variable bias. To the best of our knowledge, this is the first Healthcare Operations Management paper to leverage a quasi-randomized field experiment to mitigate omitted variable bias. We also identify a problem with the standard wait time metric, propose an alternative wait time metric and, using that alternative metric in the quasi-randomized experiment, find a somewhat smaller but significant ExWAH.

We identify two mechanisms by which a low-acuity patient delays high-acuity patients. The first is delay before triage, i.e., before a high-acuity patient is categorized as such. The second is transition delay that occurs when the ED preempts treatment of a low-acuity patient in order to treat a high-acuity patient – for example, moving a low-acuity patient out of a bed, mid-treatment, can take several minutes.

Theoretical analysis of a queueing system with transition delays in preemption reveals three ways to reduce wait times for high-acuity patients. One is process standardization to reduce the standard deviation or mean of transition delays. Second, whereas one might think that high-acuity patients benefit from having preemptive priority, eliminating preemption can reduce their wait times, under
specified conditions. Third, with or without preemption, reducing low-acuity utilization of ED resources reduces wait times for high-acuity patients.

We describe multiple ways to reduce low-acuity utilization of ED resources. One, for example, is to treat some low-acuity patients in the waiting room (without assigning them to beds required for high-acuity patient treatment). In observational data from SMMC, we show that that change in protocol substantially reduces wait times for high-acuity patients.

Methodological contributions are in §5, which develops a framework for computing the ExWAH and a robust lower bound on the ExWAH, and in §8, which uses queueing theory to characterize the average wait time for high-priority customers in a queueing system with transition delays in preemption, allowing for a general arrival and reneging process for low-priority customers.

2. Literature Review

Priority Queues in ED Research. Both preemptive and non-preemptive priority queues appear in papers modeling EDs and supporting hospital facilities, as described in (Green 2006) and citations therein. In practice, EDs follow preemptive rules for high-acuity treatment (Chisholm et al. 2000); consequently, the preemptive priority models predominate. Examples include (Lin et al. 2014), which uses a preempt-resume multi-priority $M/G/c_1/\infty$ queue to estimate patients’ wait times to access the ED; (Siddharthan et al. 1996), which models an ED as a preemptive priority queue with a Poisson arrival process and exponential service times; and (Fiems et al. 2007), which uses a preemptive priority queue to model an ED's radiology facilities. As a notable exception, (Saghafian et al. 2014) uses a nonpreemptive queueing simulation to estimate the impact of complexity-augmented triage. Other papers incorporate both preemptive and nonpreemptive customer classes when modeling EDs in order to account for a range of acuity levels, or represent preemptible and nonpreemptible stages of treatment (Gupta 2013, Laskowski et al. 2009).

Empirical ED Research. In the aforementioned preemptive-priority models of an ED, high-acuity patients wait when the system is congested with other high-acuity patients, but do not wait due to low-acuity patients. Those modeling assumptions are supported by the empirical emergency medicine literature, which concludes that high-acuity patients wait longer when an ED is crowded (McCarthy et al. 2009) and that low-acuity patients have negligible impact on high-acuity patients’ wait times (Schull et al. 2007). No further empirical emergency medicine literature tests the effect of low-acuity patients on high-acuity wait times, perhaps because the above papers conveyed the expected results.

However, the standard approach in (McCarthy et al. 2009, Schull et al. 2007) of aggregating data by time interval potentially obscures effects; moreover, (Schull et al. 2007)’s confidence intervals are small due to correlations across samples, suggesting that the small effect they do detect may
actually be 0. Our analysis in §4 uses the same wait time metric as these papers, but finds via patient-level analysis and more sophisticated causal inference that low-acuity patients do impact high-acuity wait times. In particular, we employ the general propensity score matching approach developed in (Imbens and Hirano 2004) for non-binary treatment variables to estimate the effect of the number of low-acuity patients in the ED on high-acuity wait times. Recently, (Hu et al. 2016) used a related binary treatment matching approach to study patient mortality rates and length-of-stay in the ICU.

We also employ instrumental variables derived from a quasi-randomized field experiment to estimate the ExWAH. This type of causal analysis is innovative in combining two different approaches used in recent Healthcare Operations Management literature. First, our field experiment is similar to (Song et al. 2017), which uses a treatment-control comparison between two EDs to estimate the impact of publicizing physicians’ performance metrics. Second, our use of instrumental variables falls in the same camp as (Shmueli et al. 2003, KC and Terwiesch 2012, Chan et al. 2016), which use instrumental variables to reduce the omitted variable bias in estimating a causal effect. Our instruments are guaranteed to be uncorrelated with the error terms because they are obtained from a quasi-randomized experiment.

In related literature, Batt and Terwiesch (2017) and Chan et al. (2016) show how workload influences service times in an ED and ICU, respectively, and Patterson et al. (2016) find “cherrypicking”: residents sign-up earlier for patients with simpler complaints. Freeman et al. (2016) provide an excellent survey of the broader literature on how scale and workload influence a hospital’s operational performance, and find that the length of stay in hospital for emergency patients increases with the volume of elective inpatients.

Motivation and Means for Reducing High-Acuity Wait Times. The seminal work of Naor (1969) shows that optimizing the arrival rate of customers in a service system requires quantification of the additional wait time that one arrival imposes on all other customers. Optimizing the arrival rate to an ED (“admission control”) is well-studied in Healthcare Operations Management. Three recent examples include (Xu and Chan 2016), which considers how to optimally divert incoming ED patients given future information; (Deo and Gurvich 2011), which considers how to coordinate ambulance diversion between neighboring hospitals; and (Braverman et al. 2016), which considers how to distribute patients across multiple hospitals in a network. Our work complements the ED admission control literature and (Naor 1969) by quantifying the additional wait time one low-acuity patient imposes on all high-acuity patients.

Whereas many Operations Management papers rely on ambulance diversion as the admission control mechanism, this paper instead leverages an on-site experiment in which the wait time forecast for low-acuity patients at triage influences their propensity to abandon the ED. Related
recent empirical research examines how anticipated wait time influences customer demand. In ED settings, Dong et al. (2015) find that externally published wait time information affects patients' decisions of which ED to visit, and Batt and Terwiesch (2015) find that (in the absence of a wait time forecast) the queue length, flow rate and actual wait times all impact abandonment. In call center settings, Qiu et al. (2017), Yu et al. (2017a) and Yu et al. (2017b) show how delay announcement impact customers’ waiting costs and consequently their abandonment patterns.

A rich literature gives insights into high-acuity patients’ waiting costs. (Bernstein et al. 2009) and citations therein report associations between wait time and other ED performance metrics, and can be used to estimate waiting costs. For example, for some high-acuity patients, waiting for the start of treatment can be fatal (Herlitz et al. 2005, Cardosos et al. 2011). Waiting increases patients' subsequent lengths-of-stay in the hospital (Chan et al. 2016), which drives up the cost of treatment (Kaiser 2014, Dasta et al. 2005) and increases the risk of acquiring secondary infections (Dulworth and Pyenson 2004, Donowitz et al. 1982). Additionally, waiting prolongs patients' pain, increases psychological suffering, and decreases patient satisfaction (NQMC 2016). A wide range of evidence in turn links low satisfaction with poor compliance with provider-recommended care and poor health outcomes (Doyle et al. 2013). In summary, an increase in wait times for high-acuity patients poses great costs and risks for those patients.

Variations on the Preemptive-Priority Queue. A standard preemptive-priority queueing model of an ED would not exhibit low-acuity impact on high-acuity wait times (Hassin and Haviv 2003). The empirical existence of this effect, paired with the fact that high-acuity treatment is typically preemptive, suggests that a more nuanced model of preemptive priority is needed. This critique builds on (Armony et al. 2015), which questioned whether simple queueing models could adequately characterize an ED based on exploratory data analysis in an Israeli hospital. As a first step, we discuss in §8.1 how a preemptive priority queue with transition delays from preemption could partially explain why low-acuity patients impact high-acuity wait times.

Several other papers have analyzed different variants of a preemptive-priority queue, though not in the ED context. Most notably, Cho and Un (1993) and Drekic and Stanford (2000) characterize the optimal policy in an M/G/1 queue when the decision to preempt depends on the progress of a customer’s service, and Koole (1997) characterizes the dynamic optimal policy in an M/M/1 preemptive priority queue with two priority classes and switching costs. In the above models preemption itself does not incur a delay; consequently, preemption always minimizes the wait time for high-priority customers. In contrast, our model in §8.1 identifies conditions under which a nonpreemptive policy minimizes high priority customers’ wait times.
3. Data
3.1. Observational Data
We employ observational data sets from four hospitals. Three are private teaching hospitals located in New York City that did not provide permission to be named, so are called Hospitals 1, 2 and 3. The fourth, the San Mateo Medical Center (SMMC), is a non-teaching, public hospital located in San Mateo County, California. Figure 1 shows the basic process flow at all four hospitals’ EDs, with relevant wait time metrics indicated. The data is adapted from patient visits records: for every patient who visited one of the EDs during a 15-month interval (at SMMC) or 2-year interval (at the NYC hospitals), we use timestamps for the events in Figure 1 to derive information about that patient’s visit – such as the hour of the day in which she arrived, how many other patients were in the ED when she arrived, and how long she waited for treatment. The data also includes information about a patient’s “acuity level” – need for immediate treatment due to a potentially severe medical condition – in the form of the patient’s “triage level”. A nurse first assigns the triage level during the “Triage” step shown in Figure 1. Patients with highly time-sensitive medical conditions are assigned a triage level 1 or 2, and are referred to as “high-acuity.” Patients that can wait are assigned triage levels 3, 4 or 5, and are referred to as “low-acuity.” Both high- and low-acuity patients experience dissatisfaction from waiting; however, as a matter of definition, high-acuity patients’ waiting costs are orders of magnitude larger than low-acuity patients’ waiting costs due to greater discomfort and heightened risk of adverse health events (NQMC 2016).

3.2. Experimental Data
At SMMC, in addition to the 15-month observational dataset mentioned above, we employ a dataset of 2817 ED visits that occurred during an on-site experiment we ran in Fall 2015. In §5, we limit analysis to the observational datasets; however, in §6 we use the experimental dataset to validate the analysis of §5. The remainder of this section describes the experimental design, and the resulting experimental dataset.
Beginning in February 2015 we collaborated with SMMC to publish estimates of patients’ wait times to treatment on a screen in the ED’s triage room. An automated program generates these wait time estimates using the “Q-Lasso” method, which integrates queueing theory with statistical learning to obtain wait time estimates that are much more accurate than those used by a majority of U.S. hospitals (Ang et al. 2015).

Between August 13, 2015 and October 15, 2015, we ran an experiment in which we inflated the wait times displayed in the triage room to the nearest interval of 15 minutes (for example, a Q-Lasso estimate of 14 minutes would display on the screen as 15 minutes, while a Q-Lasso estimate of 16 minutes would display on the screen as 30 minutes). Because Q-Lasso estimates the wait time based on the state of the ED, this framework introduces quasi-randomness in that patients with estimates within $\pm \epsilon$ of $15k$ for $k$ in $\{1, 2, \ldots\}$ arrive to a similar ED but see wait time estimates that have been inflated by different amounts. This approach to establish causality is similar to the regression discontinuity designs of (Lee and Lemieux 2010) and citations therein.

The experimental data contains the same basic ED visit information as the observational data, but also includes variables pertaining to the experiment – most importantly the forecast inflation seen by each patient, as deduced from the time at which he or she was triaged. §6.1 reports the results of this experiment and leverages it to develop an instrumental variable.

4. Propensity Matching Analysis
In this section, we introduce propensity score matching to improve the approach used in the emergency medicine literature to estimate the impact of low-acuity patients in the system on high-acuity patient’s wait time. We employ the standard wait time metric as in (Schull et al. 2007). However, (Schull et al. 2007)’s analysis aggregates the data across consecutive 8-hour intervals, and uses autoregressive modeling to estimate the association between the number of low-acuity arrivals in each interval and the mean and median high-acuity wait time. In contrast, we perform analysis on a properly randomized subsample of individual high-acuity patients to minimize in-sample correlations, and employ propensity score matching to reduce covariate imbalance and to estimate the causal effect of low-acuity patients on high-acuity wait times more accurately. In order to precisely state the causal inference problem, let $y_i$ be the $i^{th}$ high-acuity patient’s wait time to start of treatment, $Q_{i}^{LA}$ be the number of low-acuity patients who are waiting for start of treatment upon patient $i$’s arrival, $S_{i}^{LA}$ be the number of low-acuity patients who are in service (treatment) upon patient $i$’s arrival, and $\bar{C}_i$ be a set of control variables capturing other data

---

1 This design and timeline arose after extensive discussion with SMMC and the Institutional Review Board. We decided in particular to only make small changes to wait time estimates: the maximum adjustment of 14 minutes is less than half the standard deviation in low-acuity wait times. Per Maister’s first law of service, we also decided to inflate all estimates so as to hopefully boost patient satisfaction (Maister 2005).
about the circumstances of patient $i$'s arrival, such as the numbers of high-acuity patients awaiting treatment and in treatment, the number of doctors currently working, and indicators characterizing patient $i$'s arrival time, temperature, and weather conditions.

We adapt the continuous treatment propensity score matching approach of (Imbens and Hirano 2004) to estimate the effect of the low-acuity queue $Q_{LA}$ and low-acuity in-treatment count $S_{LA}$ on high-acuity wait times $y$ after controlling for $C$. Details regarding our adaptation are provided in the electronic companion. We bootstrap the analysis using random subsets of high-acuity patients who visited the ED over a given time period, taking one high-acuity data point per day to avoid correlations across samples. After controlling for propensity score, we find that high-acuity patients who arrived to longer low-acuity queues generally waited for significantly longer than those who arrived to shorter low-acuity queues, as shown in Figure 2. Looking at the average derivative of these curves, weighted by the number of low-acuity patients for each value of the $x$-axis, we can roughly estimate that adding one low-acuity patient to the queue increased high-acuity wait times by $1.5 \pm 0.07$ minutes at SMMC, $0.7 \pm 0.02$ minutes at Hospital 1, $0.8 \pm 0.1$ minutes at Hospital 2, and $1.1 \pm 0.1$ minutes at Hospital 3. In the case of the number of low-acuity patients in service: adding one low-acuity patient undergoing treatment increased high-acuity wait time by an average of $0.8 \pm 0.05$ minutes at SMMC, $0.14 \pm 0.01$ minutes at Hospital 1, $0.14 \pm 0.03$ minutes at Hospital 2, and $0.13 \pm 0.05$ minutes at Hospital 3.

**Figure 2** Average high-acuity wait time to treatment vs. the number of low-acuity patients waiting for treatment (top) and in treatment (bottom) upon the high-acuity patient's arrival, using data from all hospitals. The gray areas show bootstrapped 95% confidence intervals.
The primary limitation of this analysis is that, whereas Imbens and Hirano (2004) assumes that conditioned on covariates \( \bar{C} \), the outcome \( y \) and the treatment variable \( Q^{LA} \) (or \( Q^{SA} \)) are independent, that is not valid in this context because omitted variables can influence both the number of low-acuity patients and high-acuity wait times. Hence our estimates, though improved relative to (Schull et al. 2007), may still be biased. We quantify the amount of bias in our causal estimation procedure of §5 (estimating the ExWAH) via a field experiment in §6. However, we cannot use the experiment to correct the estimates of this section due to lack of enough experiment data. Specifically, units of observation in this section are high-acuity patients (that are a very small fraction of our experiment data) in contrast with §5 that relies on low-acuity patients as units of observation.

5. Estimating the ExWAH

ED managers and policy makers can influence the arrival rate of low-acuity patients in various ways (see §8 for three examples). As shown in (Naor 1969), one must account for all the extra waiting that a low-acuity arrival imposes on high-acuity patients to identify the optimal arrival rate for low-acuity patients. This section shows how to estimate that ExWAH, and is organized as follows: 5.1 defines the ExWAH; 5.2 describes the computational approach used to estimate the ExWAH empirically; 5.3 identifies how to compute standard errors and bounds for the ExWAH; 5.5 reports the resulting ExWAH estimates for each hospital using observational data.

5.1. Defining the ExWAH

The waiting externality is the difference between the aggregate waiting time of all other patients when a given patient is present and when she is not (Haviv and Ritov 1998). For example, consider a single patient, Mary. Any patient who is waiting for treatment when Mary arrives, who arrives during Mary’s visit, or who arrives between Mary’s departure and the end of the current busy period (that is, the next time the ED is idle – i.e., has zero patients in it) may have their wait time affected by Mary’s visit (Hassin and Haviv 2003).

The ExWAH is the portion of the waiting externality that affects high-acuity patients. Numerically, it is the sum of the changes in all high-acuity patients’ wait times to treatment caused by Mary’s arrival to the ED. In particular, let the “cumulative wait times of high-acuity patients potentially affected by Mary” be the sum of wait times over all high-acuity patients who were either waiting for treatment when Mary arrived, or who arrived between Mary’s arrival and the end of the current busy period. Mary’s ExWAH is the marginal change in this “cumulative wait time” caused by Mary’s arrival to the ED. When calculating it, we can restrict attention to patients who arrived to the ED in the same busy period as Mary. (We ignore correlation across busy periods, assuming it to be negligible.)
At a high level, we can estimate the ExWAH imposed by a single low-acuity patient by setting the cumulative wait time as the dependent variable in a regression, including the “number of low-acuity patients in the ED,” $N^{LA} := Q^{LA} + S^{LA}$, as a treatment variable, and fitting the regression on data from low-acuity patients. $N^{LA}$’s coefficient will then capture the average ExWAH: the change in the aggregate wait time for high-acuity patients induced by the arrival of a low-acuity patient to the ED.

We introduce additional structure to the estimation process, described in the next subsection.

5.2. Computational Approach

Suppose Mary arrived to the ED at time $t_0$, and that the busy period during which she arrived ended at time $T > t_0$.\(^2\) Let $\eta_h$ denote the sum of wait times taken over all high-acuity patients who were waiting for the start of treatment at time $t_0$, or who arrived to the ED in the interval $[t_0, \min(t_0 + h, T)]$. We call $h$ the “length of the arrival window,” and refer to such patients as “overlapping” with Mary “within a window of $h$ hours.” For every arriving low-acuity patient, we can calculate $\eta_h$ explicitly from the data by adding the wait times of all high-acuity patients who overlapped with him or her within a window of length $h$. Since $\eta_h \geq 0$ and has a long right tail, while $\log(\eta_h)$ is approximately symmetric and normally distributed for $h \geq 1$, our analysis uses the log form.\(^3\)

We will fit the regression shown in (1), where $N^{LA}$ is the number of low-acuity patients in the ED, $\bar{C}$ comprises a set of control variables (for example: the triage level of the incoming patient, the number of doctors currently working, high-acuity patient counts $Q^{HA}$ and $S^{HA}$, and time-related indicators), and $\epsilon$ is noise.

$$\log[\eta_h(Q)] = \beta_0 + \beta_1 N^{LA} + \beta_2 \bar{C} + \epsilon.$$ \hspace{1cm} (1)

Note that if we increase $N^{LA}$ by one – that is, add an additional low-acuity patient to the ED – the model (1) predicts the sum of wait times over a window of length $h$ a follows:

$$\eta_h(N^{LA} + 1) = \exp[\beta_0 + \beta_1(N^{LA} + 1) + \beta_2 \bar{C}] = \exp(\beta_1) \cdot \exp[\log(\eta_h)] = \exp(\beta_1) \cdot \eta_h(N^{LA}) .$$ \hspace{1cm} (2)

Let $X_h(N^{LA})$ denote the ExWAH truncated to a window of $h$ hours, given that $N^{LA}$ low-acuity patients are in the ED at the start of the window. Specifically, the truncated ExWAH over a window of $h$ hours is the amount by which the sum of wait times $\eta_h(N^{LA})$ increases if we increase

\(^2\) In the historical data from SMMC, the ED is empty every 11 days on average, so a busy period lasts an average of 11 days (standard deviation: 13 days). In the NYC hospital data, the system never fully empties, so $T$ last until the end of the data. For all four hospitals, we consequently approximate the ExWAH using smaller intervals, and observe that the estimates become stable as the interval length increases.

\(^3\) In the cases where $\eta_h = 0$ (up to half the time when $h = 1$; never for $h > 12$ hours), we replace $\log(0) = -\infty$ with $-5$. This corresponds to a cumulative wait time of $\eta_h = 0.04$ seconds.
$N^{LA}$ to $N^{LA} + 1$ – that is, if we introduce one additional low-acuity patient to the ED. Using (2), we formulate $X_h(N^{LA})$ as:

$$X_h(N^{LA}) = \eta_h(N^{LA} + 1) - \eta_h(N^{LA}) = e^{\beta_1} \eta_h(N^{LA}) - \eta_h(N^{LA}) = \eta_h(N^{LA})[e^{\beta_1} - 1].$$  (3)

Now, we define the ExWAH (denoted by $X(N^{LA})$) by

$$X(N^{LA}) := \lim_{h \to \infty} X_h(N^{LA}).$$  (4)

### 5.3. Getting the ExWAH from the Truncated ExWAH

**Assumption 1.** The extra waiting that Mary imposes on high-acuity patient $i$ (denoted by $\xi_i$) is nonnegative.

**Theorem 1.** If Assumption 1 holds, then the ExWAH $X$ is lower bounded by $(\max_h X_h)_+$.  

First, we show that $X_h$ is nondecreasing in $h$. Let $N(h)$ denote the set of high-acuity patients who overlapped with Mary within a window of length $h$ hours. From Assumption 1, for every $\delta > 0$,

$$X_{h+\delta} = \sum_{i \in N(h+\delta)} \xi_i = \sum_{i \in N(h)} \xi_i + \sum_{i \in N(h+\delta) \setminus N(h)} \xi_i \geq X_h.$$  (5)

Therefore, $X_h \geq \max_{h \leq h'} X_{h'}$. This means Mary’s ExWAH $X$ that is $\lim_{h \to \infty} X_h$ is lower bounded by $\max_h X_h$ (This is always nonnegative, but if empirically we obtain a value less than 0, we can use 0 as the lower bound).  

Combining (3) with Theorem 1 we can estimate the ExWAH by regression (1) for a range of values of $h$, obtaining estimates $\beta_1(h)$, and observing behavior of $\eta_h \cdot \{\exp[\hat{\beta}_1(h)] - 1\}$ as $h \to \infty$.

Furthermore, the truncated ExWAH over a window of length $h$ is bounded by $\eta_h \cdot \{\exp[\hat{\beta}_1^{LB}(h)] - 1\}$ and $\eta_h \cdot \{\exp[\hat{\beta}_1^{UB}(h)] - 1\}$, if $\hat{\beta}_1^{LB}(h)$ and $\hat{\beta}_1^{UB}(h)$ are error bars of $\beta_1(h)$. To correctly estimate $\hat{\beta}_1^{LB}(h)$ and $\hat{\beta}_1^{UB}(h)$, we incorporate several statistical adjustments that are explained below.

### 5.3.1. Clustering

We must cluster the standard errors for each $\beta_1(h)$ to account for correlation between successive patient visits. To reflect operational patterns within each ED, we cluster based on each ED’s sequencing priority schedule: in particular, during the day each ED operates a “Fast Track,” or small team of providers and nurses who focus only on treating low-acuity patients. Fast Track does not operate overnight or very early in the morning, when utilization is typically lower. Low-acuity patients who arrive when Fast Track is “on” typically wait for less time – and subsequently complete treatment more promptly – than low-acuity patients who arrive when Fast Track is “off.” We therefore divide each day into pre-Fast Track, Fast Track and post-Fast Track periods, and cluster on the resulting $3 \times 365$ groups.
5.3.2. Correction for Multiple Hypotheses Since we fit a different regression for each value of $h$, we must correct our p-values to account for multiple hypotheses. We use a standard Bonferroni correction: for each value of $h$ and corresponding estimate $\hat{\beta}_1(h)$, we define $\hat{\beta}_1^{0.001}(h)$ to be the 0.1% lower bound of $\beta_1(h)$, and $\hat{\beta}_1^{0.999}(h)$ to be the 99.9% upper bound of $\beta_1(h)$. This corresponds to a Bonferroni correction with 19 regressions, for each $h \in \{1, 2, \ldots, 12 \text{ hours}\} \cup \{1, 2, \ldots, 7 \text{ days}\}$.  

We intentionally allow the above clustering and correction approaches to be quite conservative, as doing so allows us to construct robust and reliable bounds on the ExWAH.

5.4. Explanatory Variable Selection: Defining $\bar{\mathcal{C}}$

$\bar{\mathcal{C}}$ in (1) can potentially include dozens of explanatory variables. However, including every such variable may obscure any effect attributable to $N^{LA}$, due to high correlation between $N^{LA}$ and the other covariates. To reduce $\bar{\mathcal{C}}$ to a sparser and more informative subset, we perform variable selection on a holdout dataset. At SMMC, the historical data spans one year and three months, so we used the first three months for the holdout data and one year for analysis. The NYC hospitals’ data spans 2 years each, so we used one year for holdout data and one year for analysis.

We tested using both Lasso and gradient boosting to identify good explanatory variables, but found that the latter did not significantly change the recommended variables despite being computationally more expensive (Hastie et al. 2009). Though the exact recommended variables differed by hospital, in general the pre-analysis selected triage-level indicators, the number of doctors, a subset of the arrival-time indicators, and the high-acuity patient count variables $Q^{HA}$ and $S^{HA}$.

5.5. Results

We now report the results of the above estimation process. Table 1 shows estimates of $\beta_1$, clustered standard errors, and Bonferroni-adjusted lower bounds obtained via OLS, for all four hospitals and all values of $h$. Since the truncated ExWAH over a window of length $h$ converges to the ExWAH as $h \to \infty$, these results can be thought of as lower bounds on the ExWAH.

Figure 3 shows the resulting average truncated ExWAH as a function of $h$. To generate these figures, we multiply each low-acuity patient’s $\eta_h$ by $\exp[\beta_1(h)] - 1$, $\exp[\beta_1^{0.001}(h)] - 1$ and $\exp[\beta_1^{0.999}(h)] - 1$. We then impose the requirements of Theorem 1, which states that the truncated ExWAH is nonnegative and nondecreasing in $h$. After these adjustments, the average truncated ExWAH converges to to 9.6 minutes at SMMC (mean 95% Bonferroni-corrected confidence interval: [3.2, 33.3]), 3.1 minutes at Hospital 1 (mean 95% Bonferroni-corrected confidence interval: [0.4, 9.0]), 5.9 minutes at Hospital 2 (mean 95% Bonferroni-corrected confidence interval: [5.5, 19.3]) and 6.0 minutes at Hospital 3 (mean 95% Bonferroni-corrected confidence interval: [0, 15.3]). For clarity,

$^4$ For $h > 7$ days, we observe that further extending the window $h$ does not yield additional information: effectively, the bound of Theorem 1 has already converged.
Table 1  \( \beta_1 \), clustered standard errors, and Bonferroni-corrected lower bounds for a range of window sizes of \( h \) hours.

<table>
<thead>
<tr>
<th>( h )</th>
<th>SMMC</th>
<th>Hospital 1</th>
<th>Hospital 2</th>
<th>Hospital 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \beta_1 )</td>
<td>( \beta_{1,0.01} )</td>
<td>( \beta_1 )</td>
<td>( \beta_{1,0.01} )</td>
</tr>
<tr>
<td>1</td>
<td>0.049</td>
<td>0.005</td>
<td>0.015</td>
<td>0.014</td>
</tr>
<tr>
<td>2</td>
<td>0.04</td>
<td>0.005</td>
<td>0.025</td>
<td>0.018</td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
<td>0.004</td>
<td>0.019</td>
<td>0.008</td>
</tr>
<tr>
<td>4</td>
<td>0.019</td>
<td>0.004</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>5</td>
<td>0.014</td>
<td>0.004</td>
<td>0.002</td>
<td>0.006</td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
<td>0.003</td>
<td>0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>7</td>
<td>0.007</td>
<td>0.003</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>8</td>
<td>0.007</td>
<td>0.003</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>9</td>
<td>0.011</td>
<td>0.004</td>
<td>0.003</td>
<td>0.005</td>
</tr>
<tr>
<td>10</td>
<td>0.002</td>
<td>0.003</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>11</td>
<td>0.009</td>
<td>0.004</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>12</td>
<td>-0.003</td>
<td>0.003</td>
<td>-0.002</td>
<td>-0.002</td>
</tr>
<tr>
<td>24</td>
<td>0.001</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>48</td>
<td>0.008</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>72</td>
<td>0.006</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>96</td>
<td>0.005</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>120</td>
<td>0.004</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>144</td>
<td>0.004</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>168</td>
<td>0.004</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Figure 3 also shows the average values of \( \eta_h \) multiplied by \( \exp[\beta_{1,0.025}(h)] - 1 \) and \( \exp[\beta_{1,0.975}(h)] - 1 \) (the error bars on the average truncated ExWAH obtained without a Bonferroni correction).

Figure 3  Mean ExWAH (thick solid line); mean error bars with clustered SEs and no Bonferroni correction (dotted lines); and mean lower and upper bounds with clustered SEs and Bonferroni correction (dashed lines) vs. \( h \), for SMMC, Hospital 1, Hospital 2 and Hospital 3. The x-axis (thin solid line) is drawn for clarity.

The above results might be affected by omitted variable bias between the dependent variable and \( N^{LA} \). Indeed, the next section obtains a larger estimate of the ExWAH using an experiment at SMMC, providing evidence of that bias, and §6.2 discusses examples of the omitted variables.
6. Experimental Validation

In this section, we leverage the triage room experiment described in §3.2 to show that the ExWAH is larger at SMMC than the estimate obtained in §5.5.

6.1. Experimental Results

As described in §3.2, during Fall 2015, 2817 patients at SMMC’s ED observed forecasts of their expected wait time with varying degrees of inflation. We discovered that patients who saw more inflated wait times were significantly less likely to leave without being seen (LWBS) – that is, to abandon the ED prior to their would-be start of treatment.

Specifically, let the forecast inflation \( \Delta := (\text{displayed wait time} - \text{Q-Lasso predicted wait time}) \) rounded to the nearest minute between 0 and 14 minutes. The larger the value of the forecast inflation, the more the forecast has been increased beyond the Q-Lasso estimate. Q-Lasso penalizes underestimation and overestimation equally, so a larger forecast inflation increases the chance that the number shown on the screen will exceed a patient’s actual wait time. The logistic regression (6), which controls for triage-level indicators and a set of time-related indicators \( \bar{T} \), yields a forecast inflation effect of \( \beta_1 = -0.07 \) (0.03)*. Robustness checks using probit and linear regressions yield similarly significant effects; moreover, replicating the analysis on an additional 2847 ED visits that occurred during the same time period, but during which patients did not see any wait time forecast because the screen was off, yields no significant effect \( (p = 0.98) \). (Additional details can be found in §A, including coefficient estimates in Table 3.)

\[
P(\text{LWBS}) = \beta_0 + \beta_1 \Delta + \beta_2 I\{\text{Triage Level} = 4\} + \beta_3 I\{\text{Triage Level} = 5\} + \beta_4 \cdot \bar{T} \quad (6)
\]

While SMMC’s LWBS rate is low (typically hovering between 1% and 4%), we can leverage the forecast inflation’s small but statistically significant effect to develop an instrumental variable to correct for the omitted variable bias in our estimation of ExWAH: in particular, the fact that large forecast inflations reduced the low-acuity LWBS rate means that large forecast inflations also increased the low-acuity patient counts \( Q^{LA} \) and \( S^{LA} \) beyond what they would have been otherwise. Information about recent forecast inflations is therefore a valid instrument for the number of low-acuity patients in the ED during the experimental data collection period.

6.2. Using the Experiment as an Instrument

We fit the ExWAH regression (1), but now instrument \( N^{LA} \) using the average forecast inflation taken over the 30 minutes preceding that patient’s arrival. We chose 30 minutes because it is twice as large as the maximum forecast inflation and thus covers sufficient variation, yet is not so large.

---

5 The screen intermittently turned off for a variety of reasons, most commonly due to a faulty Internet connection.
that the average forecast inflation becomes uninformative. As described in §5.3, we construct error bounds using Bonferroni-corrected, clustered standard errors. We once again cluster by daily pre-Fast Track, Fast Track and post-Fast Track intervals, which here divides our 2817 observations into 87 unique groups.

Table 2 shows estimates of $\beta_1$ obtained via OLS and 2SLS, clustered standard errors, Bonferroni-adjusted lower bounds, and $p$-values for 2SLS weak instruments and Wu-Hausman diagnostic tests for all values of $h$. Notably, our forecast inflation instruments appear to be valid: we reject weak instruments and reject the Wu-Hausman consistency test, suggesting that OLS is inconsistent.

Figure 4 shows the resulting average ExWAH as a function of $h$. The average ExWAH converges to 54.8 minutes (mean 95% Bonferroni-corrected confidence interval: [11.7, 102.9]) at SMMC. Recall that the corresponding range of ExWAH reported in §5.5 was just [3.2, 33.3] minutes. We primarily attribute this difference to omitted variable bias.

From a queueing theoretic perspective, this is unsurprising: given limited resources (staff; beds), any attempt the ED makes to increase the high-acuity service rate will decrease the low-acuity service rate, and vice versa. Each of these service rates is in turn inversely related to the high-acuity wait time and to the number of low-acuity patients in the ED. Omitting any variable with opposing effects on these two metrics will deflate the effect size when regressing one on the other. An example of such an omitted variable is information about anticipated ambulance arrivals: when the ED receives notice that one or more high-acuity patients is en-route via ambulance, it focuses resources on preparing the ED for those patients’ arrivals, thereby allowing the incoming high-acuity patients to begin treatment immediately upon arrival at the ED. This allocation of ED resources toward the incoming high-acuity patients delays treatment and discharge for other, low-acuity patients, thereby increasing the number of low-acuity patients in the ED.
In addition to omitted variables bias, two other factors may drive the difference in results between §5.5 and this section: first, we calculated the results reported in §5.5 using data from 2013, and this section uses data from 2015; perhaps the ExWAH increased over the intervening years. To test this hypothesis, we restricted the analysis of §5.5 to time periods corresponding to the experimental data and obtained an average ExWAH of just 15.2 minutes (mean 95% Bonferroni-corrected confidence interval: [5.3, 30.2]). This suggests that only a small portion of the difference in the ExWAH estimate is attributed to the time shift. Second, an instrumental variables analysis inherently emphasizes the data where the instrument has an effect, which could inflate its estimates. In particular, if the low-acuity patients that are impacted by the instrument are systematically different (have a higher ExWAH) than all low-acuity patients then it could explain the difference in ExWAH estimates.

**Figure 4** Mean ExWAH (thick solid line); error bars with clustered SEs and no Bonferroni correction (dotted lines); and mean error bars with clustered SEs and Bonferroni correction (dashed lines) vs. h at SMMC, using experimental data (n = 2817). The x-axis (thin solid line) is drawn for clarity.

### 7. What Are Causes of the ExWAH?

This section provides evidence of three different mechanisms that each contribute to the ExWAH as estimated above: delay before triage; delay in provider sign-up (an issue with wait time measurement rather than actual wait time); and transition delay when a high-acuity patient preempts a low-acuity patient.

#### 7.1. Delay before Triage

The ED does not know some high-acuity patients’ acuity levels prior to triage and in such cases triage is not preemptive. Hence triage of such high-acuity patients can be delayed by low-acuity patients.

At SMMC, the data suggests that some portion of extra high-acuity waiting indeed stems from triage delays, but that triage delays do not explain the remainder of the ExWAH, which is substantial: as shown in Figure 5, replicating the 2SLS analysis using post-triage wait times yields an average maximum truncated ExWAH of 44.9 minutes (mean 95% Bonferroni-corrected confidence interval: [15.2, 87.2]).
interval: [9.1, 85.4]) at SMMC. This is smaller than the truncated ExWAH shown in Figure 4, which converged to 54.8 minutes. (The other hospitals’ data does not include triage timestamps, so we cannot replicate this analysis in their cases.)

Figure 5 Mean post-triage ExWAH (thick solid line); mean error bars with clustered SEs and no Bonferroni correction (dotted lines); and mean error bars with clustered SEs and Bonferroni correction (dashed lines) vs. \( h \) at SMMC, using experimental data (\( n = 2817 \)). The x-axis (thin solid line) is drawn for clarity.

7.2. Delay in Provider Sign-up

The standard metric for a patient’s wait time, used in the empirical emergency medicine literature surveyed in §2, is the time elapsed between a patient’s registration and when a provider signs-up to treat the patient. That metric is widely interpreted and publicized as the wait time to “see” a provider and start treatment (which we document and discuss in the last paragraph of §9 and footnote 7). However, for some patients, we observe delay in sign-up, i.e., the provider starts to treat the patient before the provider signs-up to do so. Specifically, a typical first step in treatment is for the provider to place an order (for a specific sort of nursing care, screening for an infection or other tests, administering medication, etc.) and we observe that in some cases, the provider places an order before signing-up to treat the patient. Therefore we consider an alternative wait time metric: time elapsed between a patient’s registration and the minimum of the time that the provider places an order and time that the provider signs-up to treat the patient. For many high-acuity patients, no order is placed by the provider so we must use that minimum (rather than consider order placement times alone) to calculate a wait time for each high-acuity patient. Whereas in a teaching hospital Patterson et al. (2016) observe delay in sign-up associated narrowly with trauma cases, delay in sign-up is somewhat more widespread at SMMC, occurring with some triage level 1 and 2 patients and not restricted to ambulance arrivals.

We have order placement times only in the experimental data set for SMMC so we replicate only the experimental ExWAH analysis in §6.2 with the alternative metric. The result is a mean
ExWAH of 45.4 minutes (mean 95% Bonferroni-corrected confidence interval: [7.9, 86.2]). That is significant, though somewhat shorter than the ExWAH of 54.8 minutes obtained in §6.2.

§9 explains the potential for hospitals to better measure wait times and the ExWAH by implementing real-time location tracking systems. (The four hospitals that collaborated in this study do not yet have such systems.)

7.3. Transition Delays from Preemption

Nurses at SMMC conjecture that the primary mechanism by which low-acuity patients increase wait times for high-acuity patients is transition delay in preemption, i.e., the time required to transition resources (staff or beds) from low- to high-acuity patients. Preemption of low-acuity treatment does not occur instantaneously. Some work is required to interrupt the treatment of a low-acuity patient and prepare to treat a high-acuity patient, which typically takes at least a few minutes. For example, to move a patient out of a bed mid-treatment and prepare the bed for a different patient takes at least a few minutes.

The data suggests that preemptions and the associated transition delays occur frequently: in over 25% of high-acuity visits, the provider signed up to treat a low-acuity patient less than 5 minutes before signing up to treat the high-acuity patient. Similarly, in the experimental data 27% of high-acuity patients arrived at a time when all beds were occupied, some by low-acuity patients.

Through analysis of a stylized queueing model, §8.1 shows how low-acuity utilization and the transition delays associated with preemption contribute to high-acuity waiting.

8. Queueing Analysis & Managerial Recommendations

This section develops recommendations to help managers and policymakers mitigate the impact of low-acuity patients on high-acuity wait times.

8.1. A Queueing System wherein Preemption Imposes a Transition Delay

This subsection proposes a stylized queueing model that features a transition delay associated with preemption (highlighted by ED staff, as described in §7.3) and the following additional characteristics of a hospital ED (observed in the data from all four hospitals). First, high-acuity patients often preempt the treatment of low-acuity patients. Second, the high-acuity arrival rate is remarkably stationary, in contrast to the large diurnal fluctuations in low acuity arrivals (see Figure 2 in (Ang et al. 2015)). Third, whereas a fraction of low-acuity patients leave while waiting to begin treatment, low-acuity patients rarely leave during treatment and high-acuity patients rarely leave without treatment. Fourth, the order in which low-acuity patients are served is not FIFO; it often depends upon their service time requirements.

Consider a two-class single server queue. Class 1 customers arrive according to a stationary Poisson process with rate $\lambda_1$, whereas the class 2 arrival process is general. Each customer’s service
time requirement is an independent random variable and, within each class, customers’ service times are identically distributed. Let $S_j$ be a random variable representing the service time for a customer of class $j \in \{1, 2\}$, with mean $\mathbb{E}[S_j]$ and standard deviation $\sigma[S_j]$. Class 1 customers have preemptive priority. When a class 1 customer arrives and preempts a class 2 customer in service, a transition delay occurs between the interruption of service for the class 2 customer and the start of service for the class 1 customer. Let $D$ be a random variable representing that delay, which has mean $\mathbb{E}[D]$ and standard deviation $\sigma[D]$. Service is preempt-resume in that an interrupted class 2 customer needs only to complete her residual service time. When all class 1 customers have been served, the class 2 customers are served; the server is never idle while customers await service. The order in which class 2 customers are served is general, and in particular may depend on their service time and residual service time requirements. Whereas a class 2 customer may renege while waiting for her service to begin, after starting service, a class 2 customer stays in the system until the service is completed. In contrast, all arriving class 1 customers are served, first in first out.

Theorem 2 provides an expression for class 1 customers’ limiting average wait time ($w_1 \equiv \lim_{n \to \infty} \sum_{i=1}^{n} W_1^{(i)}/n$ with $W_1^{(i)}$ denoting the wait time to start service for the $i^{th}$ class 1 customer) in terms of the long run average fraction of time that a class 2 customer occupies the server $\rho_2$. We assume that $w_1$ and $\rho_2$ exist, $w_1$ is finite and $\rho_2 > 0$.\(^6\) Theorem 2 characterizes $w_1$ in the status quo with preemption and, for comparison, without preemption. Preemption may influence system utilization by class 2 customers (by influencing their wait times and hence reneging) so we add subscript $p$ or $p$ to $\rho_2$ to denote preemption versus no preemption, respectively.

**Theorem 2.** In the status quo with preemptive priority, the limiting average wait time of class 1 customers is, w. p. 1,

$$
\frac{\lambda_1(\mathbb{E}[S_1] + \sigma[S_1]^2)/2 + \rho_2p\mathbb{E}[D] + \rho_2\lambda_1(\mathbb{E}[D]^2 + \sigma[D]^2)/2}{1 - \lambda_1\mathbb{E}[S_1]}.
$$

(7)

With non-preemptive priority, the limiting average wait time of class 1 customers is, w. p. 1,

$$
\frac{\lambda_1(\mathbb{E}[S_1] + \sigma[S_1]^2)/2 + \rho_2n\mathbb{E}[S_2] + \rho_2\lambda_1(\mathbb{E}[S_2]^2 + \sigma[S_2]^2)/2}{1 - \lambda_1\mathbb{E}[S_1]}.
$$

(8)

The proof of Theorem 2, which employs PASTA and a generalization of Little’s law, is in Appendix B. Though (8) is known for an M/G/1 non-preemptive priority queue (Example 10-1 in (Wolff 1989)), proof is provided to establish (8) for this more general setting.

Assuming $\rho_2p = \rho_2n$, Theorem 2 implies that preemption strictly increases class 1 customers’ limiting average wait time if and only if

$$
\mathbb{E}[S_2](1 + \frac{\sigma[S_2]^2}{\mathbb{E}[S_2]^2})/2 < \mathbb{E}[D] + \lambda_1(\mathbb{E}[D]^2 + \sigma[D]^2)/2.
$$

(9)

\(^6\) Theoretically oriented readers should note that Theorem 2 only requires this assumption to hold with probability 1. Moreover, the limits $w_1$ and $\rho_2$ may be sample path dependent. The proof of Theorem 2 operates on sample paths.
One should interpret the left side of (9) as the expected residual service time for a class 2 customer whose service is interrupted by a class 1 arrival. That would be the expected wait for the arriving class 1 customer without preemption. Even if the expected transition delay in preemption \( E[D] \) is strictly lower (preemption immediately benefits the arriving class 1 customer), practiced over time, preemption results in higher average wait times for class 2 customers— if the arrival rate of class 1 customers or the expectation or standard deviation of the transition delay is sufficiently large. The rationale is that class 1 customers wait longer due to transition delays caused by previous arrivals of other high acuity patients.

Theorem 2 suggests three insights of importance for an ED. First, whereas one might have thought that high-acuity patients benefit from having preemptive-priority, Theorem 2 shows that preemption can increase high-acuity patients’ wait times. Preemption tends to increase high-acuity patients’ wait times when the high-acuity arrival rate is large, the transition delay has a large standard deviation or mean, and the expected residual service time for the low-acuity patient is small. Relatedly, Fast Track (having a subset of physicians in the ED focus on low-acuity patients) might benefit high-acuity patients. By serving low-acuity patients without preemption and associated transition delays, Fast Track reduces the effective workload for an ED and might thereby spare a high-acuity patient a transition delay.

Second, Theorem 2 shows the need to scrutinize and standardize the operating processes associated with transition delays, in order to reduce either the standard deviation or mean of the transition delays, and thereby to reduce wait times for high-acuity patients. Practical and effective ways to do so are suggested in §8.5.

Third, wait times for high-acuity patients increase with low-acuity utilization of ED resources, whether or not high-acuity patients have preemptive priority. That insight motivates the subsequent subsections §8.2-§8.5, which describe various ways to decrease low-acuity utilization of ED resources.

### 8.2. Low-Acuity Admission Control via Bed Allocation

For a hospital like SMMC where beds are a very scare resource, beds – rather than personnel – best mimic the role of a server in a queueing model. In particular, a change in protocol to treat some low-acuity patients without putting them into a bed is a form of admission control, and effectively reduces the low-acuity arrival rate. Consequently, our first recommendation concerns bed allocation rules for triage level 3 low-acuity patients.

Triage level 3 patients are assigned that triage level because they are of low-acuity and require more resources than the other (level 4 and level 5) low-acuity patients (Tanabe et al. 2004). Under standard operating procedure, Level-3 patients thus spend substantially longer in beds than Level-4
and Level-5 patients: at SMMC, for example, Level-3 patients spend an average of 3 hours (SD: 2.2 hours) between the start of treatment and discharge, compared to 1.5 hours (SD: 56 minutes) for Level-4 patients and 1 hour (SD: 41 minutes) for Level-5 patients. Level 3 patients also comprise a plurality of SMMC ED visitors, at 47% of incoming patients.

Fortunately, many level-3 patients do not require a bed in order to begin treatment. Therefore, in April 2016 SMMC began piloting a new program called “Fast Task” (distinct from “Fast Track”), under which providers perform an initial examination and begin blood tests for Level-3 patients who have waited for at least 30 minutes, without assigning the patient to a bed. While tests are running, the nurse returns the patient to the waiting area to await results, diagnosis and further treatment. This intervention eliminates the need for some Level-3 patients to be in a bed at all, and reduces time spent in a bed for others.

At SMMC, we compare high-acuity wait times during June-July 2015 (pre-Fast Task) and June-July 2016 (post-Fast Task) to estimate the benefits of this policy for high-acuity patients. (This data has no overlap with the datasets used for our other analyses.) Most notably, the mean high-acuity wait time decreased by 26% after the ED launched Fast Task, from 25 minutes to 18 minutes. While this is not a randomized comparison, the decrease persists when controlling for time-of-day, day-of-week, the number of patients in the ED, the queue length, and patients’ triage levels. Moreover, the gains appear to arise from the reduction in the effective low-acuity arrival rate, as actual arrival rates did not significantly change between the two time periods. The data also suggests that Fast Task’s reduction in the effective low-acuity arrival rate is on average equivalent to adding an additional bed to the ED: the average bed occupancy rate (number of occupied beds divided by 17 total beds) in June-July 2015 was 0.74, or about 13 beds out of 17; in June-July 2016, it decreased to 0.68, or about 12 beds out of 17. These improvements suggest that Fast Task may be a good option for other bed-constrained EDs.

8.3. Low-Acuity Admission Control via Public Policy

A demonstrably effective way to decrease the low-acuity arrival rate to EDs is to enact public policies, such as the Affordable Care Act, that provide patients with better access to primary care (Hernandez-Boussard et al. 2014). Indeed, extrapolating from (Antwi et al. 2015), we estimate that the decrease in low-acuity arrivals induced by the ACA – approximately 1.6 fewer ED visits per 1000 population per quarter – led to about 4800 fewer low-acuity ED visits in San Mateo County each year. Combining this with the ExWAH and with results of (Cardosos et al. 2011, Singer et al. 2011, Kaiser 2014) and (Aloe et al. 2009) about the impacts of waiting on patient mortality, a back-of-the-envelope calculation suggests that the decrease in the ExWAH alone saved at least 14 lives at SMMC in 2014.
8.4. Low-Acuity Admission Control via Wait Time Quotation

Many patients arriving in the ED do not truly require emergency treatment: some would be better served by primary care, and others’ conditions would resolve without treatment, as evidenced by the fact that 78% of low-acuity patients who leave without being seen do not revisit the ED within 10 days. In light of this fact and the ExWAH, it may be socially optimal that such patients LWBS after triage.

In particular, Dong et al. (2015) show that long wait times published externally (such as on a hospital’s website) can decrease the arrival rate of patients to an ED by shifting those patients to other hospitals. On the other hand, the experiment described in §3.2 showed that longer wait times displayed internally (such as on a screen in the triage room) can decrease the LWBS rate.

Some hospitals have begun to offer low-acuity patients inside the ED the option to have telemedical treatment with a short wait, whereby they may be seen by a physician located at a different hospital, and wait much less than for standard treatment (Reddy 2017).

In deciding whether, where, and how to display wait time information to current and prospective low-acuity patients, managers should account for the costs and benefits of such a policy for high-acuity patients.

8.5. Efficiency Improvement

Other papers and focus groups have proposed relevant ways to improve ED efficiency. For example, a 2007 report by the Institute of Medicine suggested that EDs should appoint special teams to accelerate bed turnover between patients (IOM 2007). (Song et al. 2017) showed that publicizing performance feedback can improve efficiency in the ED, by motivating ED staff to identify and adopt “best practices” of the staff members with relatively strong performance. Through public performance feedback, EDs could potentially reduce low-acuity patients’ service time requirements and utilization of ED resources. Similarly, EDs could potentially reduce the mean and standard deviation of transition delays in preemption through public performance feedback. However, such performance feedback would require measurement of low-acuity utilization of ED resources and transition delays in preemption- which might be accomplished through technological innovation discussed in the next section.

9. Concluding Remarks

Because high-acuity patients have preemptive priority in an emergency department, conventional wisdom holds that high-acuity patients are not caused to wait by low-acuity patients. To the contrary, this paper provides empirical evidence from four hospitals that low-acuity patients increase high-acuity patients’ wait time to start treatment. Our propensity matching analysis shows that a high-acuity patient’s wait time increases significantly with the number of low-acuity patients
who are in the system when he arrives, in observational data from four hospitals. Taking Naor's (1969) perspective, our regression analysis focuses on the total increase in wait times for high-acuity patients caused by arrival of one low-acuity patient, and finds that this is significant, in observational data from four hospitals. To rule out that these findings could be entirely due to omitted variable biases in our estimates, we undertake a quasi-randomized experiment at one of the four hospitals, and find that an arriving low-acuity patient on average increases the aggregated wait times of high-acuity patients by 55 minutes.

That effect size in the quasi-randomized experiment is much larger than in the observational results, demonstrating the importance of experimentation for causal analysis in healthcare. In contrast to the surge of randomized experiments occurring in digital advertising, e-commerce, and the social sciences, most causal analyses in healthcare are performed using observational data, perhaps due to a perception that experimentation in healthcare would be unethical or costly (Sibbald 1998). To the contrary, this paper shows that a small quasi-randomization in the wait time estimate provided at triage for low-acuity patients (a low-cost experiment posing little or no risk for patients) can correct a large bias in an observational study.

We identified two mechanisms by which low-acuity patients cause high-acuity patients to wait. The first is delay prior to triage, i.e., delay that occurs before patients are classified as being of high- or low- acuity. The second is transition delay in preemption, when service of a low-acuity patient is interrupted for a high-acuity patient.

Through queueing theoretic analysis of a stylized model with transition delays in preemption, we identified three ways to reduce wait times for high-acuity patients. The first is to reduce the standard deviation or mean of transition delays in preemption. The second is to avoid preemption, under specified conditions (the high-acuity arrival rate is large, the expected residual service time for the low-acuity patient is low, or the transition delay has a high standard deviation or mean). The third is to reduce utilization of ED resources by low-acuity patients.

We highlight three ways in which ED managers and policy makers could potentially reduce ED utilization by low-acuity patients and thereby reduce wait times for high-acuity patients. The first is to strategically begin tests and other treatment for some low-acuity patients in the waiting room, without bed assignment, to keep beds open for high-acuity patients. The second is to enact public policy to decrease unnecessary visits to the ED, e.g., by improving primary care access. The third is to allow patients that don’t require emergency treatment to choose to leave without being seen, after triage, and is related to the design of systems for providing wait time information. The triage room experiment demonstrates that providing wait time information can influence ED utilization by low-acuity patients; in ongoing research for a companion paper, we are studying how ED patients respond to wait time information.
Our results also suggest two important uses for a real-time location tracking system. Technology now exists for tracking the movements of patients and staff in an emergency department in real time, but has not yet been widely adopted. One important use would be to measure transition delays that occur when a high-acuity patient preempts treatment of a low-acuity patient, and identify the specific staff member(s) associated with each such delay. For example, the tracking system could capture the time elapsed as specific staff move a low-acuity patient out of a bed, prepare the bed, and install a high-acuity patient. We recommend public reporting of the standard deviation and mean of transition delays for each staff member. As documented in Song et al. (2017), such public relative performance reporting leads to identification and widespread adoption of best practices, which could reduce the standard deviation or mean of transition delays in preemption, and thus reduce wait times for high-acuity patients.

The second important use for a real-time location tracking system is to measure wait times. The emergency medicine literature uses the elapsed time from registration to provider sign-up as the standard metric for a patient’s wait time, and the U.S. Center for Medicare and Medicaid Services provides ED wait time information to the public and financial incentives for hospitals based on that standard metric.\(^7\) However, that metric is problematic. We observe that in some cases the provider places an order before signing-up for a patient (indicating that sign-up is delayed until after the start of treatment). Therefore, we consider an alternative metric for wait time: elapsed time from registration until the minimum of the order-placement and sign-up times, and find that an arriving low-acuity patient on average increases the aggregated wait times of high-acuity patients by 45 minutes; that effect is significant though smaller than the 55 minute effect size with the standard metric. A hospital with a real-time location tracking system could better measure wait times, apply the method proposed in this paper to estimate the waiting for high-acuity patients caused by a low-acuity patient, and thereby better measure the effect size. More generally, we encourage hospitals to implement real-time location tracking systems to look for discrepancies between “reality” and time-stamp data in electronic medical records and the managerial implications.

Appendix A: Additional Details: Experiment Results (§6.1)

We fit (6) on data from the 2817 low-acuity patients who arrived while the triage room screen was fully operational. The treatment variable \( \Delta \) denotes the forecast inflation and \( \tilde{T} \) indicates a set of time related variables. We selected these control variables following analysis on historical data (before the experimentation period) of which variables are most predictive of a low-acuity patient leaving without being seen. Table

\(^7\) See https://projects.propublica.org/emergency/ and https://www.medicare.gov/hospitalcompare/. For a typical hospital and for the 4 hospitals studied in this paper, the “time patients spent in the emergency department before they were seen by a healthcare professional” reported on these websites is based on the provider sign-up time in a patient’s electronic medical record, though some hospitals report shorter wait times by having a physician at triage.
Appendix B: Proofs for §8.1

We introduce the following random variables:

- \( t_j^{(i)} \), \( W_j^{(i)} \), and \( S_j^{(i)} \) are (respectively) the arrival time, wait time before start of service, and service time of the \( i^{th} \) class \( j \) customer (\( i \in \{1, 2, \ldots\} \) and \( j \in \{1, 2\} \)).
- \( R_{S_j}(t) \) is the residual service time for the class \( j \) customer being served, at time \( t \). If no class \( j \) customer is being served at time \( t \) then \( R_{S_j}(t) = 0 \).
- Similarly, \( R_D(t) \) is the residual transition delay if the server is in transition at time \( t \). If the server is not in transition at time \( t \), then \( R_D(t) = 0 \).
- If a class 1 customer arrives at time \( t \) and a class 2 customer is in service, the amount of transition delay faced by the class 1 customer is denoted \( D(t) \). Otherwise, \( D(t) \) is equal to 0. Note that \( R_D(t) + D(t) \) captures the transition delay of a class 1 customer that arrives at time \( t \). The two terms \( R_D(t) \) and \( D(t) \) cover mutually exclusive events. \( R_D(t) \) is positive only when the server is in transition. \( D(t) \) is positive only when a class 2 customer is served.
- \( Q_j(t) \) is the set of class \( j \) customers in queue at time \( t \), excluding the customer in service.
- \( V_j(t) \) is the total class \( j \) work in the system at time \( t \), i.e., \( V_j(t) = R_{S_j}(t) + \sum_{i \in Q_j(t)} S_j^{(i)} \).

Note: In most of the following analysis \( j = 1 \). Therefore, to simplify the notation, we drop the the subscript \( j \) whenever \( j = 1 \).

B.1. Proof of Theorem 2

Our goal is to find an expression for the long run average wait time of class 1 customers, \( w \equiv \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} W_j^{(i)} \). First, we note that

\[
W_j^{(i)} = V(t_j^{(i)}) + R_D(t_j^{(i)}) + D(t_j^{(i)}).
\]

Therefore,

\[
w = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} V(t_j^{(i)}) + \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} R_D(t_j^{(i)}) + \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} D(t_j^{(i)}),
\]

if each limit on the right hand side exists. For each in turn, we will show that the limit exist w. p. 1 and derive an expression for it.

Class 1 customers arrive according to a Poisson process, regarding which the system has no anticipation, so Theorem 7 on page 295 of (Wolff 1989) (PASTA) implies that w.p. 1,

\[
\lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} V(t_j^{(i)}) = \lim_{t \to \infty} t^{-1} \int_0^t V(u)du \text{ and } \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} R_D(t_j^{(i)}) = \lim_{t \to \infty} t^{-1} \int_0^t R_D(u)du,
\]

Table 3: Estimate for coefficient of the treatment variable \( \Delta \) in regression (6).

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Logit</th>
<th>Probit</th>
<th>OLS</th>
<th>Control (Screen Off)</th>
<th>Logit</th>
<th>Probit</th>
<th>OLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecast Inflation (( \Delta ))</td>
<td>-0.07* (0.03)</td>
<td>-0.03* (0.01)</td>
<td>-0.001* (0.001)</td>
<td>0 (0.08)</td>
<td>-0.01 (0.03)</td>
<td>0 (0.001)</td>
<td></td>
</tr>
</tbody>
</table>
if each time average limit exists and is finite. Though Theorem 7 on page 295 of (Wolff 1989) applies to indicator functions, those can be used to approximate any measurable function such as $R_D(t)$ or $V(t)$.

Theorem 1 of (Heyman and Stidham 1980) (a generalization of Little’s law) implies that w.p. 1

\[
\lim_{t \to \infty} t^{-1} \int_0^t V(u) du = \lambda_1 E[S_1]w + \lambda_1 E[S_1^2]/2.
\]  

(13)

Remark 1. Example 5-21 on page 291 of (Wolff 1989) applies that theorem to to obtain (13) for a single server queue with an arbitrary arrival process and service rule. We can map our two-class queue into that setting by completely ignoring details associated with class 2 customers or transition delays, given that Theorem 1 of Heyman and Stidham (1980) allows for an arbitrary service rule and wait time for class 1 customers. In other words, when a class 1 customer’s wait time is impacted by a transition delay we can just assume that the customer’s wait before service is impacted by an arbitrary random fluctuation. Theorem 1 of (Heyman and Stidham 1980) applies to every sample path $\omega$ of the probability space and requires only these assumptions: (a) once a job starts service it is not interrupted, (b) service times $S^{(i)}$ are independent of wait times $W^{(i)}$ for each $i$ as well as the arrival process, (c) service times $S^{(i)}$ are i.i.d, (d) $\lim_{n \to \infty} n/t^{(n)} = \lambda_i$ for sample path $\omega$, and (e) as $n \to \infty$, $n^{-1} \sum_{i=1}^n G_i$ converges to a finite limit for sample path $\omega$, where $G_i = S^{(i)}W^{(i)} + [S^{(i)}]^2/2$. In light of the assumptions we made in §8.1 and Example 5-21 on page 291 of (Wolff 1989), conditions (a)-(e) hold w. p. 1.

Next, we state a lemma for the time average of $R_D(t)$, $\mathbb{E}[R_D(t) > 0]$, and the customer average of $D^{(i)}$. Its proof is in §B.2.

**Lemma 1.** With the above definitions, we have, w. p. 1,

(a) $\lim_{n \to \infty} t^{-1} \int_0^t R_D(u) du = \lambda_1 \rho_{2p} E[D^2]/2$,

(b) $\lim_{n \to \infty} n^{-1} \sum_{i=1}^n D(t^{(i)}) = \rho_{2p} E[D]$, and

(c) $\lim_{t \to \infty} t^{-1} \int_0^t I[R_D(u) > 0] du = \lambda_1 \rho_{2p} E[D]$.

Combining (11), (13), and parts (a)-(b) of Lemma 1, we obtain

\[
w = \lambda_1 E[S_1]w + \lambda_1 E[S_1^2]/2 + \rho_{2p} E[D] + \lambda_1 \rho_{2p} E[D^2]/2.
\]

Solving for $w$ gives,

\[
w = \frac{\lambda_1 E[S_1^2]}{2} + \rho_{2p} E[D] + \lambda_1 \rho_{2p} \frac{E[D^2]}{2} = \frac{\lambda_1 E[S_1^2] + \sigma^2 E[S_1^2]}{2} + \rho_{2p} E[D] + \lambda_1 \rho_{2p} \frac{E[D^2] + \sigma^2 E[D^2]}{2}.
\]

This proves (7).

Next, we prove (8). First, recall that the subscript $j$ serves to differentiate between class 1 and class 2 customers. In the non-preemptive setting, $W_1^{(i)} = R_{S_2}(t_1^{(i)}) + V_1(t_1^{(i)})$, and therefore $w_1 = \lim_{n \to \infty} n^{-1} \sum_{i=1}^n R_{S_2}(t_1^{(i)}) + \lim_{n \to \infty} n^{-1} V_1(t_1^{(i)})$, if each limit on the right hand side exists. For each in turn, we will show that the limit exists w. p. 1 and derive an expression for it. PASTA implies that w. p. 1,

\[
\lim_{n \to \infty} n^{-1} \sum_{i=1}^n R_{S_2}(t_1^{(i)}) = \lim_{t \to \infty} t^{-1} \int_0^t R_{S_2}(u) du \quad \text{and} \quad \lim_{n \to \infty} n^{-1} \sum_{i=1}^n V_1(t_1^{(i)}) = \lim_{t \to \infty} t^{-1} \int_0^t V_1(u) du,
\]
if the time average limits exist. Given that (13) holds for an arbitrary service rule, it applies here as well (the extra wait of class 1 for class 2 customers in service can be assumed to be part of their waiting requirement). Thus, w. p. 1,
\[
\lim_{t \to \infty} t^{-1} \int_0^t V_1(u)du = \lambda_1 \mathbb{E}[S_1]w + \lambda_1 \mathbb{E}[S_1^2]/2. \tag{14}
\]
For, \( \lim_{t \to \infty} t^{-1} \int_0^t R_{S_2}(u)du \), we can use the following variant of Lemma 1(a).

**Lemma 2.** With probability 1, \( \lim_{t \to \infty} t^{-1} \int_0^t R_{S_2}(u)du = \rho_{2n} \frac{\mathbb{E}[S_2]}{\mathbb{E}[S_2]^2} \).

Combining (14) and Lemma 2, we obtain \( w_1 = \lambda_1 \mathbb{E}[S_1]w_1 + \lambda_1 \mathbb{E}[S_1^2]/2 + \rho_{2n} \frac{\mathbb{E}[S_2]}{\mathbb{E}[S_2]^2} \), that gives,
\[
w_1 = \frac{\lambda_1 \mathbb{E}[S_1^2]/2 + \rho_{2n} \frac{\mathbb{E}[S_2]}{\mathbb{E}[S_2]^2}}{1 - \lambda_1 \mathbb{E}[S_1]} = \frac{\lambda_1 (\mathbb{E}[S_1]^2 + \sigma_1^2)/2 + \rho_{2n} \frac{\mathbb{E}[S_2]}{\mathbb{E}[S_2]^2} (\mathbb{E}[S_2] + \sigma_2^2)/2}{1 - \lambda_1 \mathbb{E}[S_1]}.
\]

**B.2. Proof of Lemma 1**

Parts (a) and (b) of Lemma 1 are established by application of PASTA and Theorem 1 of (Heyman and Stidham 1980).

Denote the subset of class 1 customers that arrive while a class 2 customer is in service by \( I_{1 \to 2} \subset \{1,2,\ldots\} \). Observe that \( D(t^{(i)}) > 0 \) if and only if \( i \in I_{1 \to 2} \). Furthermore, the time average of \( \mathbb{I}[D(t) > 0] \) equals the fraction of time that a class 2 customer is in service. Using PASTA and our assumptions on \( \rho_{2n} \), the long run average fraction of time that a class 2 customer is in service,
\[
\lim_{n \to \infty} \sum_{i=1}^n \mathbb{I}[i \in I_{1 \to 2}]/n = \lim_{t \to \infty} \int_0^t \mathbb{I}[D(u) > 0]/t du = \rho_{2n} \quad \text{w. p. } 1. \tag{15}
\]

For \( i = 1,2,\ldots \) define the function
\[
f_i(t) = \begin{cases} t - t^{(i)} & \text{if } t^{(i)} \leq t < t^{(i)} + D(t^{(i)}) \\ 0 & \text{otherwise}. \end{cases}
\]
Hence, \( G_i = \int_0^\infty f_i(t)dt = D(t^{(i)})^2/2 \). Since each \( D(t^{(i)}) \) is an independent copy of transition delay, we can use SLLN and (15) to obtain, w. p. 1,
\[
\bar{G} \equiv \lim_{n \to \infty} \sum_{i=1}^n G_i = \lim_{n \to \infty} \left( \frac{\sum_{i=1}^n \mathbb{I}[i \in I_{1 \to 2}]}{n} \times \frac{\sum_{i=1}^n D(t^{(i)})^2/2}{\sum_{i=1}^n \mathbb{I}[i \in I_{1 \to 2}]} \right) = \rho_{2n} \mathbb{E}[D^2]/2. \tag{16}
\]
Note that, \( \sum_{i=1}^n \mathbb{I}[i \in I_{1 \to 2}] \) is a random variable but we can use SLLN as stated in proof of equation (14) in page 58 of (Wolff 1989). Therefore, condition (e) from Remark 1 is satisfied. On the other hand, for \( t \geq 0 \), \( H(t) = \sum_{i=1}^\infty f_i(t) \) is exactly equal to \( R_D(t) \). Therefore,
\[
\bar{H} \equiv \lim_{t \to \infty} t^{-1} \int_0^t H(u)du = \lim_{t \to \infty} t^{-1} \int_0^t R_D(u)du, \tag{17}
\]
assuming the limits in (17) exist.

Combining (16), (17), and Theorem 1 of (Heyman and Stidham 1980), we obtain that, w. p. 1, the limit \( \bar{H} \) in (17) exist and is equal to \( \lambda_1 \bar{G} \) which proves part (a) of Lemma 1.

The proof of part (b) is similar to the steps in (16), i.e., w. p. 1,
\[
\lim_{n \to \infty} \sum_{i=1}^n \frac{D(t^{(i)})}{n} = \lim_{n \to \infty} \left( \frac{\sum_{i=1}^n \mathbb{I}[i \in I_{1 \to 2}]}{n} \times \frac{\sum_{i=1}^n D(t^{(i)})}{\sum_{i=1}^n \mathbb{I}[i \in I_{1 \to 2}]} \right) = \rho_{2n} \mathbb{E}[D].
\]

Finally, the proof of (c) is similar to (a) by redefining \( f_i(t) \) to be the indicator function on \( [t^{(i)}, t^{(i)} + D(t^{(i)})] \). This gives, w. p. 1, \( \bar{G} = \rho_{2n} \mathbb{E}[D] \), and \( \bar{H} = \lim_{t \to \infty} \int_0^t \mathbb{I}[R_D(u) > 0]/du/t \). Therefore, the identity \( \bar{H} = \lambda_1 \bar{G} \) holds w. p. 1, proving (c). \( \square \)
B.3. Proof of Lemma 2

This time we apply Theorem 1 of Heyman and Stidham (1980) to the set of class 2 customers that are served. First recall that we have assumed that w.p. 1, \( \rho_{2n} \equiv \lim_{t \to \infty} t^{-1} \int_0^t \mathbb{P}[R_{S_2}(u) > 0] \, du \) exists on each sample path and is strictly positive (it may be sample path dependent). Therefore, defining \( N_2(t) \) to be the number of class 2 customers that start service up to time \( t \), we must have \( \lim_{t \to \infty} N_2(t) = \infty \) w. p. 1 because \( \rho_{2n} > 0 \) and \( \{S_2^i\}_{i \geq 1} \) is an i.i.d. sequence of random variables with a finite expectation. On the other hand,

\[
\lim_{t \to \infty} t^{-1} \int_0^t \mathbb{P}[R_{S_2}(u) > 0] \, du = \lim_{t \to \infty} t^{-1} \left[ \sum_{i=1}^{N_2(t)} S_2^{(i)} - R_{S_2}(t) \right] = \lim_{t \to \infty} t^{-1} \sum_{i=1}^{N_2(t)} S_2^{(i)},
\]

where the last step follows from the fact that \( R_{S_2}(t) \leq S_2^i \) for some \( i \) and that \( \{S_2^i\}_{i \geq 1} \) is an i.i.d. sequence of random variables with a finite expectation. Writing the right hand side as \( \lim_{t \to \infty} \left[ \sum_{i=1}^{N_2(t)} S_2^{(i)} \right] / N_2(t) \) and applying SLLN establishes that w.p. 1 \( \lim_{t \to \infty} \sum_{i=1}^{N_2(t)} S_2^{(i)} / N_2(t) = \mathbb{E}[S_2] \) so \( \lim_{t \to \infty} t^{-1} N_2(t) \) exists and is equal to \( \rho_{2n}/\mathbb{E}[S_2] \), which is a finite, strictly positive number. This fulfills one of the conditions of Heyman and Stidham (1980).

Next, if \( \tilde{t}_2^{(i)} \) is the start time of service for the \( i \)-th class 2 customer, we define the function

\[
f_i(t) = \begin{cases} t - \tilde{t}_2^{(i)} & \text{if } \tilde{t}_2^{(i)} \leq t < \tilde{t}_2^{(i)} + S_2^{(i)}, \\ 0 & \text{otherwise.} \end{cases}
\]

Hence, \( G_i = \int_0^\infty f_i(t) \, dt = (S_2^{(i)})^2/2 \) and we can use SLLN to obtain, w. p. 1,

\[
\bar{G} \equiv \lim_{n \to \infty} n^{-1} \sum_{i=1}^n G_i = \lim_{n \to \infty} n^{-1} \sum_{i=1}^n (S_2^{(i)})^2/2 = \mathbb{E}[S_2^2]/2. \tag{18}
\]

On the other hand, for \( t \geq 0 \), \( H(t) = \sum_{i=1}^\infty f_i(t) \) is exactly equal to \( R_{S_2}(t) \). Therefore,

\[
\bar{H} \equiv \lim_{t \to \infty} t^{-1} \int_0^t H(u) \, du = \lim_{t \to \infty} t^{-1} \int_0^t R_{S_2}(u) \, du. \tag{19}
\]

Finally, we note that since \( S_2^{(i)} \) are i.i.d. with finite extemation the “technical assumption” required to use Theorem 1 of Heyman and Stidham (1980) is also satisfied. Combining that result with (18) and (19), we have \( \bar{H} = (\rho_{2n}/\mathbb{E}[S_2])\bar{G} \) w. p. 1. That finishes the proof. \( \square \)

References


