

Low-Acuity Patients Delay High-Acuity Patients in EDs

Danqi Luo

Operations, Information and Technology, Graduate School of Business, Stanford University, Stanford, California 94305,
dluo@stanford.edu

Mohsen Bayati

Operations, Information and Technology, Graduate School of Business, Stanford University, Stanford, California 94305,
bayati@stanford.edu

Erica L. Plambeck

Operations, Information and Technology, Graduate School of Business, Stanford University, Stanford, California 94305,
elp@stanford.edu

Michael Aratow

San Mateo Medical Center, MAratow@smcgov.org

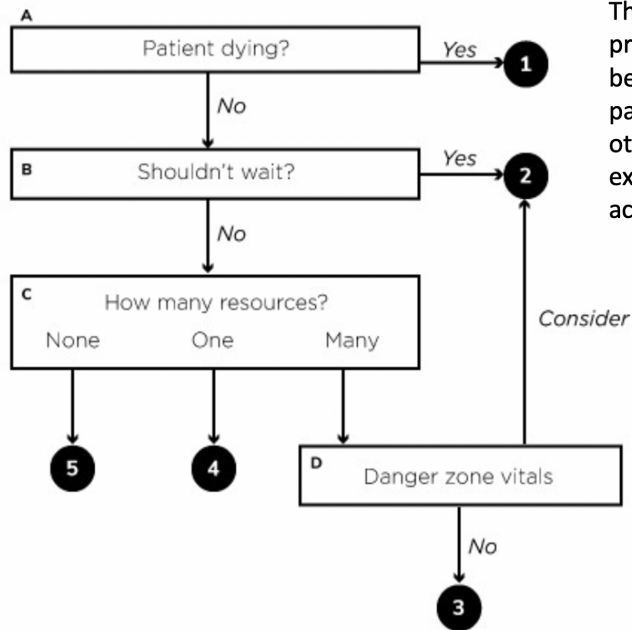
This paper provides evidence that the arrival of an additional low-acuity patient *substantially* increases the wait time to start of treatment for high-acuity patients, contradicting the long-standing prior conclusion in the medical literature that the effect is “*negligible*.” Whereas the medical literature underestimates the effect by neglecting how delay propagates in a queuing system, this paper develops and validates a new estimation method based on queuing theory, machine learning and causal inference. Wait time information displayed to low-acuity patients provides a quasi-randomized instrumental variable. This paper shows that a low-acuity patient increases wait times for high-acuity patients through: pre-triage delay; delay of lab tests ordered for high-acuity patients; and transition delay when an ED interrupts treatment of a low-acuity patient in order to treat a high-acuity patient. Hence high-acuity patients’ wait times could be reduced by: reducing the standard deviation or mean of those transition delays, particularly in bed-changeover; providing vertical or “fast track” treatment for more low-acuity patients, especially ESI 3 patients; standardizing providers’ test-ordering for low-acuity patients; and designing wait time information systems to divert (especially when the ED is highly congested) low-acuity patients that do not need ED treatment.

Key words: Health Care Management, Causal Inference, Empirical Research, Queueing Theory, Randomized Experiments

1. Introduction

Emergency Departments (EDs) operate according to the principle that high-acuity patients (HAPs) are treated before low-acuity patients (LAPs) because one or two minutes of additional *wait time* (time elapsed from arrival in the ED to start of treatment) for a HAP can increase the risk of adverse health outcomes, length of stay in hospital and associated costs. Hence operations management literature commonly models an ED as a preemptive-priority queuing system wherein LAPs have zero effect on HAPs’ wait time. A long-standing conclusion in the empirical medical literature is that LAPs have “negligible” effect on HAPs’ wait time (Schull et al. 2007, Zane 2007).

From the Handbook for the Emergency Severity Index: A Triage Tool for Emergency Department Care (Gillboy et al 2020)



The purpose of triage in the ED is to identify and prioritize the incoming patients who should not wait to be seen. A triage nurse identifies those *high-acuity* patients (ESI level 1 or 2) who should not wait. For the other, *low-acuity*, patients the nurse also evaluates expected resource needs and assigns ESI level 3, 4, or 5, accordingly.

**From the New England Journal of Medicine
Summary and Comment for Schull et al. 2007
(Zane 2007)**

“In a retrospective study of 4.2 million visits to 110 EDs, investigators assessed whether low-acuity patients affect the time to treatment for higher-acuity patients. Low-acuity patients had a clinically negligible effect on time to treatment for higher-acuity patients.”

This paper contradicts that long-standing conclusion. §5 and §7 show that the prior approach in the medical literature systematically underestimates the effect of an additional LAP on wait time for HAPs by neglecting how delays propagate in a queuing system. §4 proposes a new estimation approach; §7 validates the new approach through simulation, and shows that it corrects a substantial underestimation inherent in the prior approach. Using the new approach and observational data from five hospitals’ EDs, §4.3 shows that the effect of an additional LAP on wait time for HAPs is substantial- more than an order of magnitude greater than estimated via the prior method. Moreover, for our partner hospital, San Mateo Medical Center (SMMC), §5 incorporates a quasi-randomized instrumental variable (wait time information displayed to LAPs) to correct for unobserved variable bias, which *more than doubles* the estimated effect size. We conclude that an additional LAP has a substantial effect on HAPs’ wait time.

Through empirical and queueing theoretic analyses in §4.3-5 and §6, we characterize three mechanisms by which a LAP increases HAP wait time: (1) Delay before triage, i.e., before a HAP is prioritized; (2) Delay in lab tests for HAPs; (3) Transition delay when an ED interrupts treatment of a LAP in order to treat an HAP. In a queueing model with such transition delays, §6 shows that increasing the LAP arrival rate more greatly increases the expected wait time for HAPs when the ED is busy and crowded. Variability in transition delays and service times also exacerbates the impact of LAPs on HAP waiting. This sensitivity analysis guides the design of our estimator by indicating important interaction terms to include.

§8 describes a variety of ways to reduce wait time for HAPs, including vertical (in chairs rather than beds) treatment of some ESI 3 LAPs as piloted in the San Mateo Medical Center ED.

2. Related Literature

Carlin and Park (1970) were the first to empirically estimate the *externality*- total additional waiting for other users- caused by arrival of one additional user to a queuing system. Considering the queue of planes waiting to land at an airport, they observed that arrival of an additional plane "shoves those following it one space back in the queue for a runway" and thereby causes a delay for every following plane that propagates until the end of the busy period, i.e., until the system empties. Hence to maximize the social welfare generated by a queuing system, one must quantify the externality and deter users from waiting for service when the externality exceeds the user's private benefit, e.g., by charging a fee that reflects the externality (Naor 1969, Carlin and Park 1970). A rich literature quantifies the externality in various models of queueing systems and shows how to maximize social welfare accordingly, through sequencing priority rules, pricing, subscription, forecasting, admission control/diversion and providing wait time information; see, for example, Mendelson and Whang (1990), Haviv and Ritov (1998), Plambeck and Wang (2013), Xu and Chan (2016), Haviv and Oz (2016), Hassin and Haviv (2003) and literature surveyed therein.

The validity of standard causal inference methods, such as the method of Schull et al. (2007), relies on a "no interference" condition (Imbens and Rubin 2015) that is violated by delay propagation in a busy queuing system such as an ED. To the best of our knowledge, this paper is the first to provide a validated estimator for the expected externality for HAPs - total additional waiting for HAPs- caused by the arrival of one additional LAP to an ED.

A rich literature provides motivation for reducing HAPs' wait to start treatment in the ED. For some HAPs, one or two additional minutes of waiting for treatment can be fatal (Herlitz et al. 2005, Cardosos et al. 2011). Additional minutes of waiting increase HAPs' subsequent lengths-of-stay in the hospital (Chan et al. 2016), which in turn 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). Additional minutes of waiting prolong pain, increase psychological suffering, and decrease patients' 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). All patients suffer from waiting, but the costs and risks of waiting are far greatest for HAPs, as a matter of definition of their high-acuity triage category (NQMC 2016).

Crowding in an ED increases HAPs' wait time, associated mortality, and other adverse clinical outcomes (McCarthy et al. 2009, Bernstein et al. 2009); Hoot and Aronsky (2008) and Morley et al. (2018) survey the literature on causes and consequences of ED crowding and potential solutions.

To develop operational insight to mitigate crowding and waiting, researchers model the ED as a *priority* queuing system (HAPs are sequenced for treatment before LAPs). Xu and Chan (2016) recommend using information about future arrivals to decide when to divert LAPs to primary or urgent-care facilities and divert HAPs to another ED. Baron et al. (2019) consider joint policies for ambulance diversion and/or reservation of beds and other resources in case HAPs arrive (keeping beds empty while LAPs wait for a bed). Huang et al. (2015) characterize an asymptotically (in heavy traffic) optimal sequencing rule for doctors to provide initial and ongoing treatments to heterogeneous ED patients, to minimize congestion costs and satisfy constraints on HAPs' wait times for initial treatment. Saghaian et al. (2012, 2014) study segmentation of patients for treatment by separate teams of providers, and provide an extensive review of literature on triage, sequencing and segmentation in EDs.

EDs are commonly modeled as *preemptive* priority queues because, as explained in Chisholm et al. (2000) and Green (2006), EDs interrupt treatment of a LAP when necessary to treat an HAP. For example, Lin et al. (2014) uses a preempt-resume multi-priority $M/G/c_1/\infty$ queue to estimate patients' wait time to access the ED; Siddharthan et al. (1996) model an ED as a preemptive priority queue with a Poisson arrival process and exponential service times; and Fiems et al. (2007) use a preemptive priority queue to model an ED's radiology facilities. In these models, LAPs have no effect on the wait time for HAPs. Other papers incorporate both preemptive and non-preemptive customer classes to account for a range of acuity levels and represent preemptible and nonpreemptible stages of treatment (Gupta 2013, Laskowski et al. 2009).

The empirical healthcare operations literature shows that interrupting a provider's workflow decreases efficiency and quality of care. Gurvich et al. (2020) survey that literature and report that interrupting a physician's charting to switch to another task requires substantial changeover time.

This paper describes how interrupting treatment of a LAP in order to treat a HAP causes transition delays in an ED, and contributes to the queuing literature by analyzing a preemptive-priority queue with such transition delays. In the most closely-related queuing literature, Cho and Un (1993) and Drekić and Stanford (2000) characterize the optimal policy in an $M/G/1$ queue for a decision about whether and when to preempt, depending on the progress of a customer's service. Koole (1997) characterizes the optimal dynamic policy in an $M/M/1$ preemptive priority queue with two priority classes and switching costs.

Transition delays caused by interrupting treatment of a LAP in order to treat a HAP may help to explain empirical observations (Ardagh et al. 2002, Arya et al. 2013, Soremekun et al. 2014) that Fast Tracking LAPs (dedicating some providers to serve only LAPs, without interruption) does not increase the wait time for HAPs. Whereas most EDs Fast Track only the LAPs with minimal treatment requirements (ESI 4 and 5), Arya et al. (2013) and Soremekun et al. (2014)

suggest extending Fast Track service to ESI 3 LAPs, especially ones that could thus be treated “vertically” without occupying a bed. This paper provides empirical evidence that doing so could potentially reduce waiting for HAPs.

The empirical medical literature argues that LAPs should *not* be diverted from EDs, because they have *negligible* effect on HAPs wait times. An important nuance in this literature arises from the fact that the ESI and the Canadian Triage and Acuity Scale (CTAS) differ for LAPs: Whereas ESI levels 3, 4, and 5 sort LAPs by descending treatment resource requirements, CTAS levels 3, 4 and 5 sort LAPs by descending acuity. Vertesi (2004) argues that patients should not be sent home based upon the CTAS triage level alone, for two reasons. The first is that a fraction of patients at even the lowest-acuity triage levels CTAS 4 and 5 need ED and hospital treatment, even though they are able to safely wait for that treatment. Second, the other CTAS 4 and 5 patients require minimal ED resources for treatment, so if they were sent home without treatment, HAPs’ wait times would be “essentially unaffected”. Reinforcing that second reason, Schull et al. (2007) focus on CTAS level 4 and 5 patients who did not arrive by ambulance and were discharged home and conclude that these LAPs have a “negligible” effect on wait times for HAPs. Though Schull et al. examined the effect only of a subset of LAPs, the New England Journal of Medicine disseminated their conclusion as: “Low-acuity patients had a clinically negligible effect on time to treatment for higher-acuity patients” (Zane 2007). To the best of our knowledge, this is the first empirical paper to challenge that long-standing conclusion, perhaps because the analysis of Schull et al. (2007) is based on an impressive data set from 110 EDs. Unlike Vertesi (2004) and Schull et al. (2007), this paper considers all LAPs, and shows that an additional ESI 4 or 5 LAP causes substantial delay for HAPs (Table EC.4) and the effect of an ESI 3 LAP is greater.

The instrumental variable (IV) approach described in this paper is motivated by empirical service operations management literature. Chan et al. (2016) and KC and Terwiesch (2012) use ICU congestion as an IV to estimate, respectively, the effect of patients’ wait for admission to the ICU on their length of stay in the ICU, and the effect of patients’ length of stay in the ICU on their likelihood of readmission. We similarly use wait time information displayed to LAPs as an IV to estimate the effect of an additional LAP that waits to start treatment on HAPs’ wait time for high-acuity patients. Wait time information provided to customers influences whether or not they choose to wait for service, as demonstrated by empirical studies in call centers (Qiu et al. 2017, Yu et al. 2017a,b), a ride sharing system (Yu et al. 2020), and an ED (Ang et al. 2015). In EDs that do not provide wait time information, patients may nevertheless infer wait time information from the number of patients waiting, the flow rate, and the actual wait time, and accordingly decide whether or not to wait for treatment (Batt and Terwiesch 2015). ED wait time information published on the internet influences patients’ choice of which ED to visit (Dong et al. 2019).

Medical literature calls for reducing the number of patients that Leave an ED Without Being Seen (LWBS) (Carter et al. 2014). Relatedly, we find that inflating the wait time displayed to LAPs reduces the number that LWBS. The drawback is that, especially when the ED is busy and crowded, reducing the number of LAPs that LWBS causes HAPs to wait longer.

3. Data

We employ data from five hospitals. Hospitals 1, 2 and 3 are private teaching hospitals located in New York City and Hospital 4 is a private teaching hospital located in California. The fifth hospital, San Mateo Medical Center (SMMC), is a non-teaching, public hospital located in San Mateo County, California. Figure 1 shows the process flow at all five hospitals' EDs. For each patient visit, we have timestamps for events in squares in Figure 1 and use the timestamps to derive information such as the hour of the day in which a patient arrived, how many other patients were in the ED when that patient arrived, and the patient's wait time, defined as the time elapsed between registration in the ED and provider sign-up to start treating the patient. The data includes the

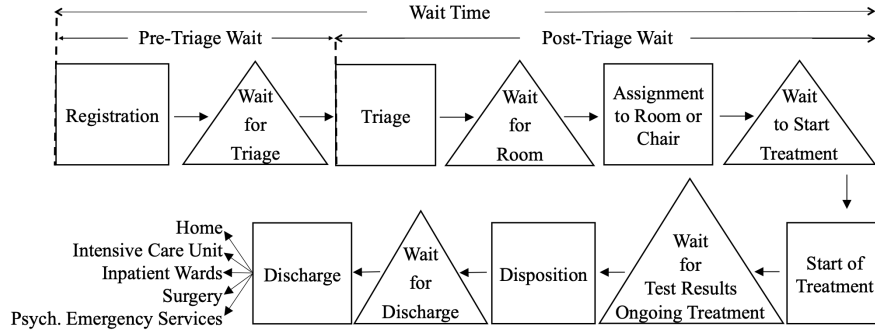


Figure 1 ED process flow diagram. Squares are events recorded in patient-visit data.

ESI index assigned to each patient at triage. Recall from page 1 that ESI 1 and 2 are the HAPs who should not wait, whereas ESI 3, 4 and 5 are LAPs.

Table 1 summarizes the time period, total number of patient visits, high- and low-acuity mean wait time, and high-and low-acuity mean arrival count in the data from each hospital. For SMMC we have data from three distinct time periods. The 2015 SMMC data is for the time period in which we partnered with SMMC to provide quasi-randomized wait time information to LAPs at triage, which serves as an instrumental variable in §5. Therefore all results in the body of this paper are based on 2015 SMMC data, with 2012-13 SMMC data for variable selection. (Appendix A reports similar results in 2019 SMMC data; during 2019 SMMC provided wait time information at registration, which creates estimation issues discussed in Appendix A.)

Table 1 Time period, number of patient visits, and summary statistics

		SMMC			Hospital 1	Hospital 2	Hospital 3	Hospital 4
Start Date		21/6/12	13/8/15	1/4/19	15/11/09	1/1/09	1/1/09	20/8/11
End Date		26/9/13	15/10/15	31/12/19	31/3/12	31/12/11	31/12/11	18/8/16
Patient visits		52,167	7,057	28,879	176,497	324,076	186,833	182,555
Mean Wait Time (Minutes)	HAP	25.3	23.7	24.6	23.5	20.5	19.4	30.7
	LAP	55.6	57.8	56.7	60.8	35.1	27.0	73.5
Mean Arrival Rate (Patients/Day)	HAP	11	12	10	14	28	17	9
	LAP	102	100	95	191	268	154	91

4. Estimating the Effect of an Additional LAP on HAP Wait Time

§4.1 explains how prior medical literature under-estimates the effect of an additional LAP on the wait time for HAPs, then §4.2 proposes our estimation approach, and §4.3 reports our estimates as well as estimates derived from Schull et al. (2007) based on the observational data.

4.1. Under-Estimation in the Emergency Medicine Literature

The approach taken in the emergency medicine literature (McCarthy et al. 2009, Schull et al. 2007) aggregates data by time interval and does not directly account for the effect of an additional arrival during the time interval on crowding and wait time during subsequent time intervals. In particular, Schull et al. (2007) aggregate data into 8-hour time intervals, regress the mean wait time for HAPs arriving in an 8-hour-interval on the number N^L of LAPs that arrive during that same 8-hour-interval, and interpret the coefficient of N^L as the effect of one additional LAP arrival per 8 hours on the mean wait time for HAPs.

A fundamental insight from queueing theory is that delay caused by an additional arrival propagates throughout a busy period for the queueing system. EDs operate continuously, 24 hours per day, and rarely empty. Hence in an ED, an arriving LAP could delay the treatment of subsequently-arriving LAP or HAPs, which in turn could delay the treatment of subsequently-arriving LAP or HAPs, and so on, causing delay for subsequent arrivals for *long* afterwards.

Therefore Schull et al.’s approach underestimates the effect of a LAP arrival on HAP waiting by failing to account for HAP waiting in a subsequent 8-hour-interval. Think, for example, of a LAP that arrives at the end of an 8-hour-interval; *all* the HAP waiting caused by that LAP occurs in a subsequent 8-hour-interval. The effect of a LAP arrival on HAP waiting in subsequent time intervals violates the *no interference* requirement in the Stable Unit Treatment Value Assumption (Imbens and Rubin 2015), invalidating Schull et al.’s estimation method and all causal inference methods that rely on SUTVA (Sobel 2006). Not only do Schull et al. underestimate the effect of a LAP arrival on HAP waiting, Schull et al.’s use of standard regression confidence intervals is incorrect because interference between adjacent 8 hour intervals (due to delay propagation) increases the correlation of outcomes in these intervals. Hence we propose a new approach that accounts for the temporal causal relationship between a LAP arrival and subsequent waiting for HAPs.

4.2. Proposed Estimation Approach

Consider the waiting externality for HAPs caused by arrival of a LAP (called “Mary”). The wait time for any HAP who arrives *after* Mary may change because of Mary’s visit to the ED (Hassin and Haviv 2003). Let ξ_i denote the change in the wait time for HAP i caused by Mary’s visit. Let $N(\tau)$ denote the set of all HAPs who arrive during the next $\tau > 0$ units of time after Mary does. The truncated externality caused by Mary is

$$X_\tau = \sum_{i \in N_\tau} \xi_i. \quad (1)$$

To estimate the expected truncated externality X_τ for an additional LAP arriving at a randomly-chosen time, we propose a temporal causal framework. We associate the number of LAPs arriving within a short time interval of length $\epsilon > 0$ with wait time for HAPs during the next τ units of time. Specifically, for a given time point t , we define predictor variable N_ϵ^L to be the count of any LAPs that arrive within the time period $[t - \epsilon, t)$ and remain in the ED to start treatment. We define η_τ to be the sum of wait time taken over all HAPs who arrive during the time period $[t, t + \tau]$. We divide each day into three 8-hour intervals (12:00 AM-7:59 AM, 8 AM-3:59 PM, and 4:00 PM-11:59 PM), randomly select one time point t within each 8-hour-interval, and calculate the variables N_ϵ^L and η_τ for each of those randomly-sampled time points. Thus, each day of data contributes three samples for the regression analysis. Due to the right-skewed distribution of η_τ , we use its log transform in the regression:

$$\log(\eta_\tau) = \beta_0 + \beta_1 N_\epsilon^L + \beta_2 N_\tau^H + \beta_3 N_\tau^L + \text{Interaction Terms} + \beta_8 \vec{C}_\tau + \text{noise}, \quad (2)$$

wherein N_τ^H and N_τ^L denote the count of HAP and LAP arrivals during $[t, t + \tau]$, respectively, \vec{C}_τ is the set of control variables for $[t, t + \tau]$, and

$$\text{Interaction Terms} = \beta_4 N_\epsilon^L N_\tau^H + \beta_5 N_\epsilon^L N_\tau^L + \beta_6 N_\tau^L N_\tau^H + \beta_7 N_\epsilon^L N_\tau^L N_\tau^H. \quad (3)$$

Motivating those interaction terms, §6 proves in a queueing model of an ED that, insofar as the system is busy with high arrival rates of HAPs and LAPs, additional LAP arrivals will more greatly increase HAP waiting. For samples with $\eta_\tau = 0$, to address the $\log(0) = -\infty$ issue, we apply the two-part model approach from Section 19.3.4 of Greene (2012). The average relative change

$$A_\tau \equiv \frac{\partial}{\partial N_\epsilon^L} \mathbb{E} [\log(\eta_\tau) | N_\epsilon^L, N_\tau^H, N_\tau^L] = \beta_1 + \beta_4 N_\tau^H + \beta_5 N_\tau^L + \beta_7 N_\tau^L N_\tau^H \quad (4)$$

represents the *multiplicative factor* by which an arriving LAP increases the aggregate wait time for HAPs during the subsequent period of length τ . Our estimator for the expected truncated externality X_τ is

$$\hat{X}_\tau \equiv \frac{1}{n} \sum_{i=1}^n [\hat{\eta}_\tau(N_{\epsilon,i}^L + 1) - \eta_\tau(N_{\epsilon,i}^L)] = \frac{1}{n} \sum_{i=1}^n \eta_\tau(N_{\epsilon,i}^L) [e^{A_{\tau,i}} - 1], \quad (5)$$

where n is the number of observations in the regression analysis, $\eta_\tau(N_{\epsilon,i}^L)$ is the observed value for the i -th sample, and the estimate $\hat{A}_{\tau,i}$ is obtained by using the estimated coefficients $\hat{\beta}_0, \dots, \hat{\beta}_7$, observed values of $N_{\epsilon,i}^L, N_{\tau,i}^L, N_{\tau,i}^H$, and Eq. (4).

Dozens of variables are candidates for \vec{C} in (2) but including them all could obscure any effect of LAP arrivals N_ϵ^L due to high correlation between N_ϵ^L and other variables. To create a sparse and informative \vec{C} , we applied Lasso for variable selection on a holdout dataset. Hospital 1, 2, and 3 provided data spanning 2 years, so we use the first year for variable selection and the second year for analysis. Hospital 4 data spans 5 years, so we used the first year for variable selection and the remaining 4 years for analysis. For variable selection for SMMC, we used historical data spanning the 15 months starting in January 2013. §EC.3 lists the selected control variables \vec{C} for each hospital. Each includes the time of day and season in which the sample time t occurs, and patient counts at time t : number of LAPs waiting to begin treatment, number of LAPs in treatment, and number of HAPs in treatment.

To generate confidence intervals, we use a bootstrap technique to calculate a 95% confidence interval for the expected truncated externality at a particular τ . Moreover, as we consider multiple levels of τ (multiple regressions) we apply a Bonferroni correction to generate larger confidence intervals that reliably contain the true 95% confidence intervals for all those levels of τ . §EC.2 provides more detail and validation for these Bootstrap and Bonferroni techniques.

4.3. Results in Observational Data

For each of the five hospitals, our estimate of the expected truncated externality \hat{X}_τ initially increases with τ and then converges. For SMMC and Hospital 4, \hat{X}_τ increases over the entire range of τ from 0.5 to 12 hours, demonstrating that the effect of an additional LAP arrival on HAP waiting can persist for longer than 12 hours; see Figure 2, top panels. For Hospitals 1, 2 and 3, \hat{X}_τ converges within τ of 8 hours, as shown in Figure 5 in Appendix A.

Table 2, top row, reports the expected externality truncated at $\tau = 8$ hours, \hat{X}_8 , for each of the five hospitals. Table 2 focuses on that 8 hour time interval for purposes of comparing estimates by our approach and that of (Schull et al. 2007). Readers should keep in mind that \hat{X}_8 excludes any effect of an additional LAP arrival on wait time for HAPs that arrive more than 8 hours later and so, particularly for SMMC and Hospital 4, underestimates the total effect of the additional LAP on HAP waiting. Increasing τ from 8 hours to 12 hours increases our estimate of \hat{X}_τ by 15% to 11 minutes at SMMC, and by 38% to 3.6 minutes for Hospital 4, respectively.

Table 2, second row, reports the estimate by Schull et al.’s method for the effect of one additional LAP arrival in an 8 hour interval on the expected wait time *per HAP* that arrives in the 8 hour interval. For comparison with Schull et al.’s per HAP estimate in Table 2, we estimate the effect of

one additional LAP arrival on the expected wait time *per HAP* arriving in the next $\tau = 8$ hours, by dividing \hat{X}_τ from (5) by the number of HAPs that arrive during the interval of length τ

$$\frac{1}{n} \sum_{i=1}^n \frac{\hat{\eta}_\tau(N_{\epsilon,i}^L + 1) - \eta_\tau(N_{\epsilon,i}^L)}{N_{\tau,i}^H} = \frac{1}{n} \sum_{i=1}^n \frac{\eta_\tau(N_{\epsilon,i}^L)[e^{\hat{A}_{\tau,i}} - 1]}{N_{\tau,i}^H} \quad (6)$$

Our per HAP estimate- and also that of Schull et al. depend on the choice of focal time interval (8 hours for the results reported in Table 2). For Hospital 1, our per HAP estimate decreases from 0.6 (± 0.1) at $\tau = 2$ hours to the 0.3 (± 0.2) at $\tau = 8$ hours reported in Table 2, apparently because the effect of an additional LAP arrival on HAP wait time largely occurs within 2 hours after the LAP arrival; see Figure 5 in Appendix A. Similarly, for Hospital 2, our per HAP estimate decreases from 0.6 (± 0.1) at $\tau = 3$ hours to the 0.4 (± 0.2) at $\tau = 8$ hours in Table 2. Figure 2 shows how our per HAP estimate varies with the length of the interval τ at SMMC and Hospital 4; analogous figures for Hospitals 1, 2, 3, and SMMC (using 2019 data) are in EC.1 with per HAP estimates summarized in the ‘per HAP’ row in Table 2 and 0.3 (± 0.2) for SMMC using 2019 data.

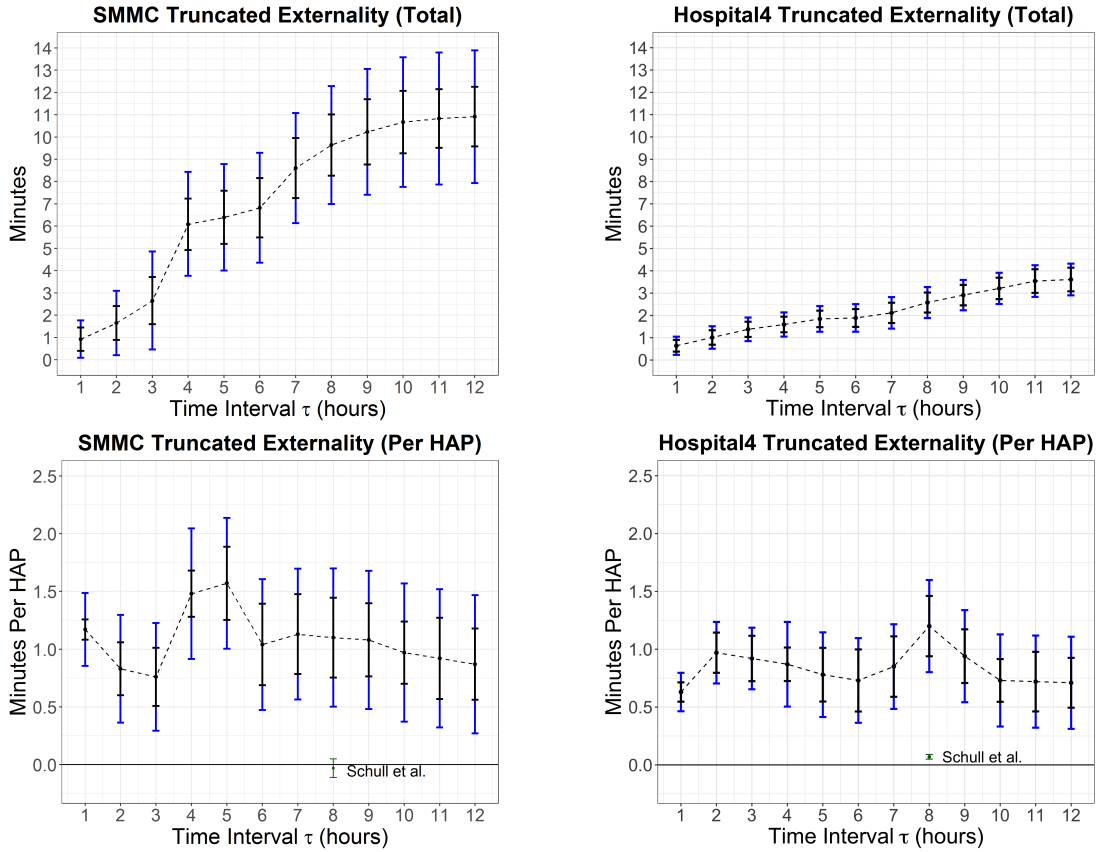


Figure 2 Estimated expected truncated externality \hat{X}_τ (dot) with bootstrap 95% confidence interval (inner error bars) and Bonferroni-corrected 95% confidence interval (outer error bars).

Table 2 Estimated effect (and 95% confidence interval) of one additional LAP arrival in an 8 hour interval on expected wait time *per HAP* and *total* for all HAPs arriving in 8 hour interval.

		SMMC	Hospital 1	Hospital 2	Hospital 3	Hospital 4
<i>total</i>	Ours	9.6(± 1.4)	2.4(± 0.6)	2.8(± 0.5)	1.9(± 0.7)	2.6(± 0.5)
<i>per HAP</i>	Ours	1.1(± 0.4)	0.3(± 0.2)	0.4(± 0.2)	0.5(± 0.2)	1.2(± 0.3)
	Schull et al.'s	-0.03(± 0.08)	-0.19(± 0.09)	-0.08(± 0.10)	0.05(± 0.02)	0.07(± 0.02)

Our estimates in Figure 2 and Table 2 show that in all five hospitals, an additional LAP arrival substantially increases the expected wait time for HAPs, contradicting (Schull et al. 2007). Schull et al.'s approach incorrectly suggests that an additional LAP reduces HAP wait time in Hospital 1 and has negligible effect in the other hospitals. (Recall from §4.1 that Schull et al.'s method of generating confidence intervals, replicated with the Schull et al. estimates in Figure 2 and Table 2, also is incorrect.) The effect is clinically significant because, as documented in §2, an additional minute or two of wait time can increase risk of mortality, adverse health outcomes and length of hospital stay for a HAP. Moreover, our estimates show the effect of *one* additional LAP arrival on HAP wait time. In Hospital 3, with smallest estimated total effect of 1.9(0.7) minutes of extra HAP wait time per LAP arrival, 268 LAPs arrive per day, so if 5 or 10% of those LAPs could be safely diverted, the benefit for HAPs could scale up accordingly.

Why is the total effect of an additional LAP arrival on expected wait time for HAPs higher at SMMC than Hospital 4, whereas the per HAP effect is similar at SMMC and Hospital 4? One reason apparent in Table 1 is that the HAP arrival rate is higher in the SMMC 2015 data than in Hospital 4. Intuitively, the effect of an additional LAP arrival on wait time for HAPs that arrive in the next τ units of time would tend to increase with the number of HAPs $N_{\tau,i}^H$ that arrive, whereas our per HAP estimate (6) decreases with that $N_{\tau,i}^H$.

To shed light on when and how LAPs delay HAPs, we study each of the three stages of waiting depicted in Figure 1: pre-triage wait (time elapsed from registration to start of triage), post-triage wait (time elapsed from start of triage to start of treatment) and wait for test results (time elapsed from when a test is ordered until the lab returns the results) shown in Figure 1. We find that LAPs substantially increase all three stages of waiting for HAPs. Furthermore, we examine how the post-triage externality differs during hours that an ED operates a Fast Track, in comparison with hours with no Fast Track, and by subcategory of LAP. Schull et al. (2007) did not consider these three stages of waiting, nor account for Fast Track, likely due to data limitations. Our partner hospital SMMC provided the additional data required for these analyses, so we focus on SMMC in the remainder of this section.

Regarding *pre-triage* delay, SMMC nurses explain that the ED is forewarned of the arrival of some HAPs so can prepare to expedite them through triage to treatment, but many other HAPs

must wait before a triage nurse determines that they are of high-acuity. LAPs ahead in the queue for triage cause those HAPs to wait longer to be triaged and prioritized. To estimate the pre-triage expected truncated externality, we calculate \hat{X}_τ using the pre-triage wait time for HAPs and letting N_ϵ^L be the count of LAPs that arrive within the ϵ interval and stay to start triage. See the left panel of Figure 3. The expected pre-triage externality truncated at τ of 12 hours is 4.5 minutes (bootstrap 95% confidence interval [3.4, 5.7]).

To estimate the *post-triage* expected truncated externality, we calculate \hat{X}_τ using the post-triage wait time for HAPs, in N_ϵ^L the count of LAPs that start triage within the ϵ interval and stay to start treatment, N_τ^L the number of LAPs that start triage during $[t, t + \tau]$ and wait to start treatment, and N_τ^H the number of HAPs that start triage during $[t, t + \tau]$. See the right panel of Figure 3. The expected post-triage externality truncated at τ of 12 hours is 5.8 minutes.

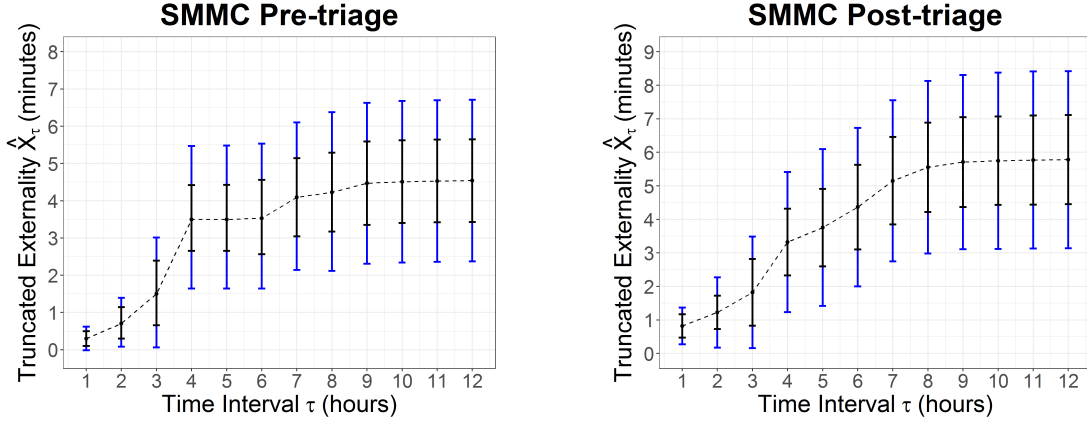


Figure 3 In left panel, the estimated expected pre-triage truncated externality \hat{X}_τ (dot) with bootstrap 95% confidence interval (inner error bars) and Bonferroni-corrected 95% confidence interval (outer error bars). In right panel, the post-triage analog.

Between 12 pm and 10 pm on weekdays, to mitigate LAPs' long wait times while the ED is most busy and crowded, SMMC operates a *Fast Track*: rather than prioritize HAPs, a dedicated set of providers treat only LAPs. This post-triage change in sequencing during “Fast Track Hours” (weekdays from 12 - 10 pm) raises the question: Is the post-triage expected truncated externality different during Fast Track Hours than at another time? To estimate the post-triage expected externality during Fast Track Hours, we adapt our estimation approach by randomly selecting t from Fast Track Hours. Similarly, we estimate the externality during “Not Fast Track Hours” by randomly selecting t from the intervals of time that are not Fast Track Hours. Having limited capacity, the Fast Track providers do not treat any ESI 3 LAPs and treat only 80% of the ESI 4 and 5 LAPs triaged during Fast Track Hours. Therefore we adapt our estimation approach to focus on categories of LAP, by ESI and whether the LAP is routed to the Fast Track.

Table 3 SMMC estimated post-triage expected externality truncated to 12 hours caused by a LAP that is triaged during Fast Track Hours versus at another time, and contingent on the category of LAP. Bootstrap 95% confidence intervals are provided in parenthesis.

	Category of LAP			
	All	ESI 3	ESI 4&5	ESI 4&5 Routed to Fast Track
Fast Track Hours	5.6(± 1.3)	6.6(± 1.5)	4.5(± 1.4)	4.1(± 1.5)
Not Fast Track Hours	5.9(± 1.4)	7.3(± 1.6)	4.6(± 1.5)	not applied

The effect of an ESI 3 LAP is statistically different - larger - than the effect of an ESI 4 or 5 LAP. A likely reason is that, by definition, an ESI 3 LAP requires more treatment time and resources than does an ESI 4 or 5 LAP.

During Fast Track Hours, the effect of one additional ESI 4 or 5 LAP routed to Fast Track may be smaller than of one not routed to Fast Track, but is not statistically different. One explanation is that Fast Track providers are capacity-constrained, so routing an additional LAP to the Fast Track would cause a subsequently-arriving LAP to be routed to the non-Fast Track providers, and thereby increase the expected wait time for HAPs.

Together, our results Table 3 and Theorem 1 in §6 suggest that perhaps Fast Track mitigates LAPs' effect on HAPs' post-triage wait times. According to Theorem 1, without a Fast Track, additional LAP arrivals would have a larger effect on HAP wait time while the ED is most busy and crowded, i.e., during Fast Track Hours. Yet the effect of an additional LAP arrival during Fast Track Hours does not statistically differ from the effect of an additional LAP arrival at another time, according to the results in the first, "All" column of Table 3. Perhaps operating the Fast Track reduces the effect of an additional LAP arrival on HAP wait time, cancelling out the increase in the effect of a LAP on HAP wait time because the ED is busy and crowded during Fast Track Hours. We conclude that a controlled trial- expanding Fast Track capacity to serve some ESI 3 LAPs - is warranted to evaluate the potential benefit for HAPs and ESI 3 LAPs.

At SMMC, HAPs and LAPs (routed to Fast Track or not) rely on the same laboratory staff and resources to process diagnostic tests. By causing a HAP to wait longer for test results, a LAP could delay the start of treatment needed by a HAP. Waiting for a test result can delay the start of treatment urgently needed by a HAP, as described in (Nørgaard and Mogensen 2012). Yet a typical HAP's wait time for a test result is not captured in that patient's wait time to provider sign-up ¹, the imperfect metric for a patient's wait time to start of treatment used elsewhere in

¹At SMMC, for 86% of HAPs, a provider first signs up to treat the patient, then orders one or more tests, then waits for test results to make decisions regarding treatment. For 12% of HAPs the triage nurse orders urgent tests, but for only 4% of those HAPs are the test results returned before the provider signs up to treat the patient.

this paper, in (Schull et al. 2007) and related literature. Therefore we estimate the effect of an additional test-order for a LAP on the expected *wait time for a test result* for a HAP, on tests for HAPs ordered within the next τ hours. We calculate \hat{X}_τ using the wait time for a test result for a HAP, in N_ϵ^L the count of orders placed for a test for a LAP within the ϵ interval, and N_τ^L and N_τ^H the number of orders placed for tests for LAPs and HAPs, respectively, during $[t, t + \tau]$. Truncated at τ of 12 hours, this is *34 minutes* (bootstrap 95% confidence interval [19, 49]). The effect remains substantial when distributed per HAP. The effect of an additional test-order for a LAP on the expected wait time per test for a HAP ordered within the next 12 hours is 5 minutes.

We conclude that an additional LAP substantially increases HAPs' expected wait time from registration to triage, wait time from triage to start of treatment, and wait time for a test result.

Moreover, any omitted variable that increases both the LAP arrival rate and HAP wait time would bias all the result in this section (for our approach and that of Schull et al.) toward under-estimation. §5 corrects for omitted variable bias in our estimate of the post-triage externality at SMMC and finds a substantially larger effect than estimated above.

5. Quasi-Randomized Instrument in Wait Time Information

At SMMC during August 13 to October 15, 2015, we displayed for 5,265 LAPs at triage their estimated remaining wait time to start treatment. We used the Q-Lasso method of Ang et al. (2015) to predict that wait time, inflated that prediction to the nearest multiple of 15 minutes, and displayed the inflated prediction on a screen in the triage room. For example, a Q-Lasso prediction of 14 minutes resulted in a display of 15 minutes, while a Q-Lasso prediction of 16 minutes resulted in a display of 30 minutes. Q-Lasso predicts the wait time based on the current state of the ED. Hence this approach introduced quasi-randomness in that patients with Q-Lasso predictions of a few minutes smaller or larger than $15k$ for k in $\{1, 2, \dots\}$ arrived to an ED in a similar state but experienced different amounts of inflation.

We measure the effect of inflation in the displayed wait time on the number of LAPs that wait to start treatment, as follows. Let Δ denote the inflation: “displayed wait time” minus “Q-Lasso predicted wait time.” SMMC updates the predicted wait time at intervals of 10 minutes. Therefore, building on the setting of (2), we choose ϵ of 10 minutes, sample t randomly from the set of time points at which an update occurs, let N_ϵ^L be the number of LAPs that arrive at triage during $(t - \epsilon, t)$ and wait to start treatment, and fit regression

$$N_\epsilon^L = \alpha_0 + \alpha_1 \Delta + \vec{\alpha}_2 \cdot \overrightarrow{\text{Controls}} + \text{noise} \quad (7)$$

for the inflation Δ in the displayed wait time during $(t - \epsilon, t)$ and the time-related control variables listed in §EC.3. By design, all the LAPs that arrive at triage during $(t - \epsilon, t)$ experience the same

inflation Δ in the displayed wait time that remains constant throughout the interval $(t - \epsilon, t)$. (Moreover, §EC.6 shows that for more than 95% of LAPs, the displayed wait time remains constant throughout their entire triage service.)

A natural experiment occurred during our data collection period of August 13 to October 15, 2015 in that, due to a faulty internet connection, the screen displaying wait time information occasionally went off. We separately fit (7) for the time periods in which the screen was On versus Off because, in the latter case, the LAPs did not see the displayed wait time information and inflation therein. The wait time inflation yields a (almost statistically significant) positive effect in each of the bootstrapped data sets when the screen is On but no significant effect when the screen is Off. For one representative bootstrapped dataset, the first row of Table 4 reports the coefficient α_1 of Δ and p-value for the two regressions. This suggests that greater inflation Δ in the displayed wait time increases the number of LAPs that wait to start treatment.

Table 4 Coefficient and p-value of inflation Δ from regression (7) on a representative bootstrapped SMMC 2015 data when Screen is On versus Off (first row). The same quantities when Δ in regression (7) is replaced with the nonlinear function of inflation $f_*(\Delta)$ generated by the Machine Learning IV algorithm (Singh et al. 2020)(second row).

	Screen On		Screen Off	
	Coefficient	p-value	Coefficient	p-value
Δ	0.03	0.12	-0.005	0.31
$f_*(\Delta)$ via MLIV	0.05	0.03	-0.004	0.15

How could greater inflation Δ in the displayed wait time increase the number of patients that wait to start treatment? Prospect theory (Tversky and Kahneman 1991) helps to explain this phenomenon. The displayed wait time is a reference point. Patients that wait longer than the reference point suffer loss and intensified disutility from ongoing waiting that discourages them waiting beyond the reference point. In contrast, waiting for less than the reference point is a gain, so patients are encouraged to wait up until the reference point. Therefore an increase in the reference point (larger inflation Δ in the displayed wait time) increases the number of patients who wait for long enough to start treatment.

Providing empirical support for that explanation, §EC.6 shows that inflation Δ in the wait time displayed to a LAP decreases the probability that the LAP leaves the ED prior to start of their treatment, i.e., reduces the LWBS rate, a statistically significant effect.

We use the method of Singh et al. (2020) to convert Δ into a strong instrument to improve our estimate of the post-triage externality. As in the regression discontinuity design of Thistlethwaite

and Campbell (1960), Δ (the amount by which our wait time prediction is inflated by rounding up to the nearest 15 minute interval) is a quasi-random instrument. Moreover, inflation Δ is not associated with the post-triage externality for HAPs except through the number of LAPs that wait to start treatment. However, the inflation Δ is a weak instrument, as is evident from the p-value of 0.12 in the first row of Table 4. This weakness arises from the small size of our dataset: it spans only 63 days and we randomly sample one t within each 8 hour interval (excluding intervals when the wait time display was down due to the faulty internet connection) so have only 142 units of observation in each bootstrapped dataset. We apply the Machine Learning IV (MLIV) algorithm (Singh et al. 2020) to construct a stronger instrument, as explained in detail in §EC.6.2. Our resulting IV is a non-linear function of the inflation, which we denote $f_*(\Delta)$. The second row of Table 4 shows that this IV has a statistically significant effect on N_ϵ^L .

We incorporate the strong IV $f_*(\Delta)$ into our regressions for the expected truncated post-triage externality in total (2) and per HAP (6). Recall that the post-triage externality is the effect of one additional LAP who arrives at triage and waits to start treatment on the post-triage wait time for HAPs that arrive at triage during the next τ hours. We fit both the regression for the total effect (2) and the regression for the per HAP effect (6) with the ϵ , t , N_ϵ^L defined exactly as for the above regression (7), with \hat{X}_τ calculated using the post-triage wait time for HAPs and with N_τ^L and N_τ^H the number of LAPs and HAPs, respectively, that start triage during $[t, t + \tau]$ respectively and wait to start treatment. Complete output of the 2SLS is reported in §EC.6.3.

Figure 4 shows the resulting estimate of the expected post-triage externality truncated to interval τ , in total and per HAP that arrives at triage during interval τ , as τ ranges from 1 to 12 hours.

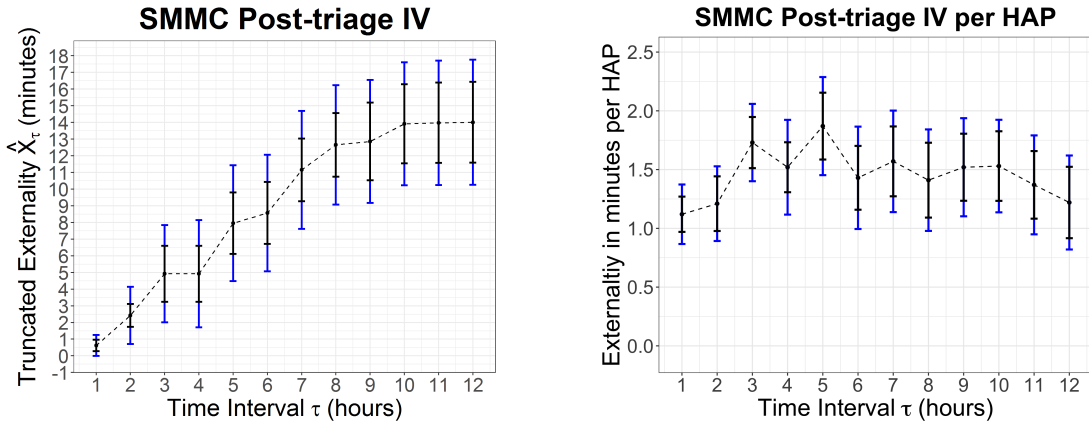


Figure 4 Mean externality (dot); mean 95% bootstrap confidence interval (inner error bars); mean 95% Bonferroni-corrected confidence interval (outer error bars).

By comparing the left panel of Figure 4 with the right panel of Figure 5, one can see that incorporating the IV more than doubles the estimated effect size. For example, truncated at τ

of 12 hours, the IV-based estimate of the expected post-triage externality is 14.0 minutes (with bootstrap 95% confidence interval of [11.5, 16.5]) whereas with only the observational data the estimate was 5.8 minutes (with bootstrap 95% confidence interval [4.5, 7.1]).

The true post-triage truncated externality may be higher than depicted in Figure 4. LAPs who are contemplating whether or not to wait to start treatment and have that decision influenced by the instrument (inflation in the displayed wait time) are presumably the LAPs least in need of services in the ED, and hence cause the least waiting for other patients. Insofar as the LAPs influenced by the instrument impose a lower externality than other LAPs, our IV approach underestimates the externality caused by a LAP.

The 2019 IV analysis (using quasi-randomized wait time information provided at registration) in Appendix A also shows that the total and per HAP effect size is substantially higher than estimated based on observational data. The estimated post-triage per HAP externality is 1.5 minutes with 95% bootstrap confidence interval [1.1, 1.9].

6. Queuing Theory

According to SMMC nurses, LAPs increase the post-triage wait time for HAPs due to the *transition delays* that occur when the ED interrupts treatment of a LAP in order to treat a HAP, i.e., the time required to transition staff, beds, or other resources from treating a LAP to treating a HAP. For example, from our direct observation at SMMC, moving a LAP out of a bed in order to put an HAP into that bed to start treatment can take more than 10 minutes. In the Toronto ED studied in Chartier et al. (2016), such a bed changeover takes more than 23 minutes on average, including 9 minutes for cleaning. When the treatment of a LAP is interrupted, the LAP typically continues to wait in the ED for ongoing treatment, and may be moved to a hallway bed or a chair.

We analyze a stylized model of an ED that features such transition delays. Consider a two-class single server queue in which LAPs and HAPs arrive by Poisson processes with rates λ_L and λ_H , respectively, and HAPs have preemptive priority, albeit with a transition delay. When a LAP is in service and an HAP arrives, a delay D occurs and then the HAP begins service. Work is conserved in that when service resumes for the LAP, only the residual service time is required. LAPs are served in FIFO order when no HAPs are in the system. LAPs may renege while waiting to begin service, but after starting service, a LAP remains in the system until the service is completed. Service times have a general distribution with mean and standard deviation of $\mathbb{E}[S_L]$ and $\sigma[S_L]$ for LAPs and of $\mathbb{E}[S_H]$ and $\sigma[S_H]$ for HAPs, respectively. Transition delays have a general distribution with mean $\mathbb{E}[D]$ and standard deviation $\sigma[D]$. Inter-arrival times, service times, and transition delays all are independent random variables.

Theorem 1 provides an expression for HAPs' limiting average wait time $w_H \equiv \lim_{n \rightarrow \infty} \sum_{i=1}^n W_H^{(i)} / n$ with $W_H^{(i)}$ denoting the wait time to start service for the i^{th} HAP. We impose

no requirement on LAP reneging, except that the resulting long run average utilization of the treatment server by LAPs ρ_L and w_H exist, $\rho_L > 0$ and w_H is finite *w.p.1.*; the limits ρ_L and w_H may be sample path dependent. Utilization ρ_L , which is the long run average fraction of time that a LAP occupies the server, may differ with and without preemption, which we indicate by appending p or n to ρ_L to denote utilization in the system with preemption vs. non-preemption.

THEOREM 1. *With preemptive priority, HAPs' limiting average wait time is, w. p. 1,*

$$\frac{\lambda_H(\mathbb{E}[S_H]^2 + \sigma[S_H]^2)/2 + \rho_{Lp}\mathbb{E}[D] + \rho_{Lp}\lambda_H(\mathbb{E}[D]^2 + \sigma[D]^2)/2}{1 - \lambda_H\mathbb{E}[S_H]}. \quad (8)$$

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

$$\frac{\lambda_H(\mathbb{E}[S_H]^2 + \sigma[S_H]^2)/2 + (\rho_{Ln}/\mathbb{E}[S_L])(\mathbb{E}[S_L]^2 + \sigma[S_L]^2)/2}{1 - \lambda_H\mathbb{E}[S_H]}. \quad (9)$$

Hence preemption strictly increases HAPs' limiting average wait time if and only if

$$\rho_{Ln} \left[\mathbb{E}[S_L] \left(1 + \frac{\sigma[S_L]^2}{\mathbb{E}[S_L]^2} \right) / 2 \right] < \rho_{Lp} \left[\mathbb{E}[D] + \lambda_H (\mathbb{E}[D]^2 + \sigma[D]^2) / 2 \right]. \quad (10)$$

The proof of Theorem 1, which employs PASTA and a generalization of Little's law, is in §B. Though (9) is known for an M/G/1 non-preemptive priority queue (Example 10-1 in (Wolff 1989)), proof is provided to establish (9) for this more general setting.

Theorem 1 shows that reducing LAPs' utilization of ED resources reduces HAP waiting. Furthermore, to the extent that an ED is busy (high utilization by LAPs ρ_L or high utilization by HAPs $\lambda_H\mathbb{E}[S_H]$) reduction in the LAP arrival rate or utilization of ED resources would more greatly reduce HAP waiting.

Moreover, reducing the mean $\mathbb{E}[D]$ or standard deviation $\sigma[D]$ of the transition delay associated with preemption reduces the effect of LAPs on HAP waiting, and thereby reduces HAP waiting. In reality, a way to reduce the mean and standard deviation of bed-changeover delays is to provide walkie-talkies to the triage nurse and staff that clean beds and move patients (Chartier et al. 2016).

Theorem 1 shows that HAP average wait time increases with both the mean $\mathbb{E}[S_H]$ and standard deviation $\sigma[S_H]$ of the time required to treat a HAP. Though our model does not explicitly incorporate testing, according to (Nørgaard and Mogensen, 2012), increasing a HAP's wait time for test results can increase the time that a provider spends treating a HAP, which would translate to an increase in $\mathbb{E}[S_H]$ and $\sigma[S_H]$. Recall from §4.3 that an additional LAP test increases the expected HAP wait time for test results. Together, these analytic and empirical insights suggest that LAPs can indirectly increase HAPs wait time to start of treatment, by increasing other HAPs' wait for test results, and thereby increasing those HAPs' mean and standard deviation in treatment time.

Theorem 1 shows that preemption increases HAPs' wait time if the arrival rate of HAPs is large or if the transition delay has a large standard deviation or large mean, relative to the expected

residual service time for a LAP (the term in brackets in the left hand side of (10)). On the other hand, preemption increases LAP waiting, which could increase LAP reneging, and thereby reduce LAP utilization to $\rho_{Lp} < \rho_{Ln}$, which would tend to reduce HAP waiting.

Is preemption possibly increasing HAP waiting at SMMC? We fit the model to match SMMC’s mean arrival rates and mean wait times for HAPs and LAPs, and LAP reneging rate of 3% with preemptive priority. We assume that with no preemption, the LAP reneging rate would drop to 0—the reneging rate that maximizes HAP waiting with no preemption. Adding the assumption that service times and transition delays are exponential, we see that preemption increases HAP mean wait time ((10) holds) when the mean transition delay exceeds 3.4 minutes, as supported by our observation and the medical literature. With a mean transition delay of 23 minutes, corresponding to the minimum bed changeover time in Chartier et al. (2016), the long run average wait for HAPs of 19.7 minutes with preemption *is reduced by eliminating preemption*, to only 6.1 minutes. This suggests that preemption might indeed be increasing HAP waiting at SMMC ED. §EC.9 reports other examples wherein preemption increases HAP waiting.

7. Validating our Estimator via Simulation

To validate our estimator, we simulate a two-class tandem queue model of an ED in which LAPs cause pre-triage waiting and post-triage waiting (transition delay in preemption) for HAPs. We extend the model formulated above to incorporate triage. Arriving HAPs and LAPs are served in First In First Out (FIFO) order on a single “triage” server, with independent exponentially distributed service times. Then they proceed to the “treatment” server with preemptive priority and transition delays as formulated above. Using simulation, we generate synthetic data and evaluate the performance of our approach and that of Schull et al. (2007) in estimating the mean truncated externality and the per HAP outcome measure proposed by Schull et al.

To do so we fit the model to observational data from five hospitals, as described below and summarized in Table 5. SMMC and Hospital 1 are similar in mean wait times, so our first simulation setup is motivated by both SMMC and Hospital 1. Arrival rates for HAPs and for LAPs are set to the average of the rates at SMMC and Hospital 1. Utilization at the triage server is set so that the mean wait for triage matches the mean in SMMC data: 8.5 minutes. Mean treatment times for HAPs and LAPs are set so as to match the mean post-triage wait times for HAPs and LAPs in SMMC data. In doing so, we assume that the transition delay D is a deterministic 5 minutes, service times are exponentially distributed, and no patients renege. We make those same assumptions in constructing each of the five parameter settings. We set utilization at the treatment server by HAPs and LAPs, respectively, to match the mean wait time for HAPs and mean wait time for LAPs at SMMC and Hospital 1. For hospitals 2, 3 and 4 we do not have triage time stamp

data, so we assume utilization at triage is the same as at SMMC. To model each of the hospitals 2, 3, and 4, we use the observational mean arrival rates for HAPs and LAPs and set the utilization at the treatment server by HAPs and by LAPs in the model ($\rho_L = \lambda_L \mathbb{E}[S_L]$ and $\rho_H = \lambda_H \mathbb{E}[S_H]$) so as to match the mean wait times for HAPs and LAPs in observational data. Our fifth parameter setting is motivated by the SMMC peak period between 10 am and noon. We match the parameter setting to the observational mean arrival rates for HAPs and LAPs, the observational mean wait time from registration to triage, and the observational mean wait time from triage to start of treatment for HAPs and LAPs in that peak period.

Table 5 Five parameter settings based on observational data from five hospitals.

		H1/SMMC	Hospital 2	Hospital 3	Hospital 4	SMMC peak
Arrival rate (patients/day)	LAP λ_L	96.0	48.0	48.0	96.0	68.6
	HAP λ_H	24.0	26.2	24.0	26.2	48.0
Utilization at Triage		0.67	0.41	0.40	0.68	0.81
Utilization at Treatment	LAP ρ_L	0.40	0.20	0.20	0.40	0.29
	HAP ρ_H	0.30	0.33	0.30	0.33	0.60
Transition Delay D (minutes)		5	5	5	5	5
Mean Wait Time (minutes)	LAP	57.2	37.4	32.5	73.4	1047.0
	HAP	25.2	20.4	19.1	27.5	51.0

To calculate the *true* truncated externality X_τ , we simulate a base scenario and a counterfactual scenario, between which the only difference is that in the counterfactual scenario, one LAP (called Mary) is randomly added to the arrival process after a burn-in period to allow the system to reach its steady-state distribution. The true truncated externality is

$$X_\tau \equiv \sum_{i \in N_H(\tau)} W_i^{\text{counterfactual}} - W_i^{\text{base}},$$

where $N_H(\tau)$ is the set of HAPs that arrive within the next τ units of time after Mary. For comparability with Schull et al.’s approach, we truncate the externality at $\tau = 8$ hours.

For each of the five model parameter settings, 200 runs of the base and counterfactual simulation were performed and the results are presented in Table 6. We report the true mean truncated externality X_τ and its 95% confidence interval in the row of “Oracle”.

For each of the five model parameter settings, we ran the base simulation for a much longer time to collect the time stamp data from 300,000 patient visits after the initial burn-in period. We use that time stamp data in exactly the same manner that we used the observational data, as reported in §4.3, to calculate our estimate of the expected truncated externality (*total* increase in wait time for HAPs in an 8 hour time interval caused by one additional LAP arrival), our estimate of the expected increase in wait time *per HAP* arriving in the 8 hour time interval, Schull et al.’s estimate

of that expected increase in wait per HAP from one additional LAP arrival per 8 hour interval, and a corresponding extrapolated estimate of the total effect from multiplying by the mean number of HAP arrivals per 8 hour interval. These estimates for the total and per HAP effect by our approach and that of Schull et al. are reported in Table 6.

The final row of Table 6, labeled “Theorem 1” reports the true increase in mean wait time per HAP caused by increasing the LAP arrival rate by one per 8 hours (the metric that Schull et al. purport to estimate). We add expression (8) in Theorem 1 for HAPs’ mean wait time for treatment to the well-known expression for mean wait time for triage (an M/M/1 queue) to obtain an expression for HAP’s total mean wait time, and how this changes as λ_L increases by one LAP per 8 hours. Details of this derivation are in §EC.7. This is a different metric than the expected truncated externality per HAP that we calculate with the Oracle simulation method and estimate by (6), which depend on the truncation interval τ .

Our estimates closely approximate the true mean truncated externality (in total and per HAP) from the Oracle simulation method. In contrast, Schull et al.’s approach underestimates the effect size. The magnitude of underestimation is greatest for the busiest system “Peak SMMC”, which makes sense in that the underestimation inherent in Schull et al is due to LAP arrivals in one time interval creating delays that propagate to increase the wait times for HAPs that arrive in subsequent time intervals within the same busy period. In the busiest system, busy periods are long-lasting, so that each LAP causes delay for more HAPs.

Table 6 Estimated effect (and 95% confidence interval) of one additional LAP arrival in an 8 hour interval on expected wait time *per HAP* and *total* for all HAPs arriving in the 8 hour interval, from the Oracle, our approach and Schull et al.’s approach. From Theorem 1, increase in mean HAP wait time from increasing the arrival rate λ_L by one LAP per 8 hours.

		H1/SMMC	Hospital 2	Hospital 3	Hospital 4	SMMC peak
<i>total</i>	Oracle	5.0(± 1.3)	3.1(± 1.0)	2.5(± 0.5)	6.7(± 1.8)	44.7(± 12.1)
	Ours	5.0(± 0.3)	3.4(± 0.2)	2.5(± 0.2)	6.1(± 0.2)	32.6(± 0.6)
<i>per HAP</i>	Oracle	1.1(± 0.2)	0.3(± 0.1)	0.3(± 0.1)	1.2(± 0.1)	7.1(± 1.4)
	Ours	0.9(± 0.1)	0.4(± 0.1)	0.3(± 0.1)	0.9(± 0.1)	6.5(± 0.5)
	Schull et al.’s	0.4(± 0.01)	0.2(± 0.01)	0.2(± 0.01)	0.5(± 0.01)	0.4(± 0.02)
	Theorem 1	1.4	0.5	0.5	1.5	6.6

We repeated this simulation analysis under the alternative assumptions that the transition delay has an exponential distribution and that the SMMC LAP and HAP arrival rates vary diurnally. Specifically, the arrival rate for each hour of the day is set equal to the mean number of arrivals during that hour of the day in the SMMC observational data. The results, reported in §EC.7, Table

EC.8, remain qualitatively the same. Our estimation approach performs well, whereas the approach of Schull et al. underestimates the effect size, especially in the busiest system “Peak SMMC”.

8. Managerial Implications

The empirical evidence that LAPs substantially increase HAPs’ wait times (to be triaged, to start treatment, and for test results) implies that ED managers can reduce HAPs’ wait times by reducing unnecessary use of ED resources by LAPs. For example, to reduce HAPs’ wait times for test results, ED managers could reduce unnecessary diagnostic testing for LAPs. Doctor et al. (2020) document large inconsistency among providers in the number of tests ordered per LAP, and recommend training providers to standardize their test-ordering and reduce unnecessary testing for LAPs. SMMC exhibits similarly large inconsistencies and opportunities to reduce unnecessary testing for LAPs. For each of the 21 MDs that work at SMMC ED during non-Fast Track hours, we computed the average number of tests for a LAP under the age of 21 treated by that MD during non-Fast Track hours between April 1, 2019, and April 1, 2020. The minimum among the 21 MDs was 1.6 tests per LAP, and the maximum was 6.3 tests per LAP. Reducing a HAP’s wait time for test results could reduce the HAP’s length of stay in a bed, reduce the amount of time a provider spends on treating the HAP, and thereby reduce other HAPs’ wait times for a bed or provider.

Another way to reduce HAP waiting would be to design wait time information systems so as to reduce the number of LAPs that choose to seek treatment in an ED when the ED is busy and crowded. Our queuing theoretic analysis shows that reducing the LAP arrival rate would more greatly reduce the expected wait time for HAPs to the extent that an ED is crowded and busy. Our experimental evidence from SMMC and literature surveyed in §2 shows that wait time information systems influence LAPs’ choices regarding whether or not to wait to be treated in an ED and which ED to visit.

Similarly, HAP waiting could be reduced by offering LAPs in the ED the option for telemedicine (consultation with a physician located elsewhere) and using wait time information systems to encourage LAPs to opt for that telemedicine when the ED is busy and crowded.

To reduce LAP bed use (and reduce HAP waiting that occurs when staff moves a LAP out of bed to put the HAP into the bed to start treatment), an ED could treat more LAPs “vertically,” in chairs rather than beds. For example, SMMC piloted “Fast Task” (distinct from Fast Track): for any ESI 3 LAP that has waited for at least 30 minutes, a provider performs an initial exam and orders tests- while the patient remains seated in a chair. Fast Task eliminates bed use by some ESI 3 LAPs and reduces the time spent in a bed for other ESI 3 LAPs. By comparing HAPs wait time during June-July 2015 (pre-Fast Task) and June-July 2016 (post-Fast Task), we find a 26% reduction in HAP mean wait time to start treatment, from 25 minutes to 18 minutes, associated

with Fast Task. This association persists when controlling for time-of-day, day-of-week, and the number of patients in the ED of each ESI level. We recommend that bed-constrained EDs conduct a randomized experiment with Fast Task or other protocol to treat more LAPS “vertically” to evaluate the effect on HAP wait time.

Whereas Fast Task reduces ESI 3 LAPs’ bed utilization and associated bed-changeover delays for HAPs, the ESI 3 LAPs would still share providers with HAPs and cause provider transition delay for HAPs. To avoid *both* bed-changeover and provider transition delays that occur when treatment of a LAP is interrupted to prioritize treatment of a HAP, an ED could expand Fast Track service to some ESI 3 patients, especially ones that can thus be treated vertically. Many EDs operate a Fast Track during the busy period of the day, but only for LAPs of ESI level 4 and 5. Among all LAPs, the ESI 3’s tend to have the longest treatment times and impose the highest waiting externality on HAPs, suggesting that expanding Fast Track service to ESI 3 LAPs could reduce HAP waiting. The caveat is that in an ED with shortages of space and providers, allocating more space and more providers to a Fast Track for LAPs might increase HAP waiting by reducing the number of providers that prioritize treatment of HAPs and their treatment space.

Beyond EDs, we hope that service operations managers and researchers will apply our approach to estimate waiting externalities in order to improve the management of other complex queuing systems. Moreover, in any service operations that provide wait time information to customers, a round-up in the estimated wait time could serve as a quasi-randomized instrument for arrivals or renegeing for purposes of causal inference, as demonstrated in this paper.

References

- Ang, E., S. Kwasnick, M. Bayati, M. Aratow, E. Plambeck. 2015. Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management* **18** 141–156.
- Ardagh, MW, J Elisabeth Wells, Katherine Cooper, Rosa Lyons, Rosemary Patterson, Paul O’Donovan. 2002. Effect of a rapid assessment clinic on the waiting time to be seen by a doctor and the time spent in the department, for patients presenting to an urban emergency department: a controlled prospective trial. *The New Zealand Medical Journal (Online)* **115**(1157).
- Arya, Rajiv, Grant Wei, Jonathan V McCoy, Jody Crane, Pamela Ohman-Strickland, Robert M Eisenstein. 2013. Decreasing length of stay in the emergency department with a split emergency severity index 3 patient flow model. *Academic Emergency Medicine* **20**(11) 1171–1179.
- Baron, Opher, Tianshu Lu, Jianfu Wang. 2019. Priority, capacity rationing, and ambulance diversion in emergency departments. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3387439 .
- Batt, R.J., C. Terwiesch. 2015. Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science* **61** (1) 39–59.

-
- Bernstein, S. L., Aronsky D., R. Duseja, S. Epstein, D. Handel, U. Hwang, M. McCarthy, K. McConnell, J. M. Pines, N. Rathlev, R. Schafermeyer, F. Zwemer, M. Schull, B. R. Asplin. 2009. The effect of emergency department crowding on clinically oriented outcomes. *Acad Emerg Med* **16**(1) 1–10.
- Burke, Paul J. 1956. The output of a queuing system. *Operations research* **4**(6) 699–704.
- Cardosos, L., C. Grion, T. Matsuo, E. Anami, I. Kauss, L. Seko, A. Bonametti. 2011. Impact of delayed admission to intensive care units on mortality of critically ill patients: a cohort study. *Critical Care* **1**.
- Carlin, Alan, Rolla Edward Park. 1970. Marginal cost pricing of airport runway capacity. *The American Economic Review* 310–319.
- Carter, Eileen J, Stephanie M Pouch, Elaine L Larson. 2014. The relationship between emergency department crowding and patient outcomes: a systematic review. *Journal of Nursing Scholarship* **46**(2) 106–115.
- Chan, C. W., V. F. Farias, G. J. Escobar. 2016. The impact of delays on service times in the intensive care unit. *Management Science* .
- Chartier, Lucas Brien, Licinia Simoes, Meredith Kuipers, Barb McGovern. 2016. Improving emergency department flow through optimized bed utilization. *BMJ Open Quality* **5**(1) u206156–w2532.
- Chisholm, C., E. Collison, D. Nelson, W. Cordell. 2000. Emergency department workplace interruptions: Are emergency physicians “interrupt-drive” and “multitasking”? *Academic Emergency Medicine* **7**(11).
- Cho, Y. Z., C. K. Un. 1993. Analysis of the m/g/1 queue under a combined preemptive/nonpreemptive priority discipline. *IEEE Transactions on Communications* **41**:1 132–141.
- Dasta, J. F., T. P. McLaughlin, S. H. Mody, C. T. Piech. 2005. Daily cost of an intensive care unit day: the contribution of mechanical ventilation. *Crit Care Med* **33**(6) 1266–71.
- Doctor, Kaynan, Kristen Breslin, James M Chamberlain, Deena Berkowitz. 2020. Practice pattern variation in test ordering for low-acuity pediatric emergency department patients. *Pediatric emergency care* .
- Dong, Jing, Elad Yom-Tov, Galit B Yom-Tov. 2019. The impact of delay announcements on hospital network coordination and waiting times. *Management Science* **65**(5) 1969–1994.
- Donowitz, L. G., R. P. Wenzel, J. W. Hoyt. 1982. High risk of hospital-acquired infection in the ICU patient. *Critical Care Medicine* **10**:6.
- Doyle, C., L. Lennox, D. Bell. 2013. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open* **3**(1).
- Drekic, S., D. A. Stanford. 2000. Threshold-based interventions to optimize performance in preemptive queueing systems. *Queueing Systems* **35** 289–315.
- Dulworth, S., B. Pyenson. 2004. Healthcare-associated infections and length of hospital stay in the medicare population. *American College of Medical Quality* **19**:3.
- Fiems, D., G. Koole, P. Nain. 2007. Waiting times of scheduled patients in the presence of emergency requests. *Working Paper* .

-
- Gilboy, Nicki, Paula Tanabe, Debbie Travers, AM Rosenau, et al. 2020. Emergency severity index (esi): a triage tool for emergency department care, version 4. *Implementation handbook* **2020** 1–17.
- Green, L. 2006. Queueing analysis in healthcare. Randolph W. Hall, ed., *Patient flow: Reducing delay in healthcare delivery*. New York: Springer-Verlag, 290.
- Greene, William H. 2012. Econometric analysis, 7e. *Stern School of Business, New York University*.
- Gupta, D. 2013. Queueing models for healthcare operations. *Handbook of Healthcare Operations Management*, vol. 184. 19–44.
- Gurvich, Itai, Kevin J O’Leary, Lu Wang, Jan A Van Mieghem. 2020. Collaboration, interruptions, and changeover times: Workflow model and empirical study of hospitalist charting. *Manufacturing & Service Operations Management* **22**(4) 754–774.
- Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Springer Science & Business Media.
- Haviv, M., Y. Ritov. 1998. Externalities, tangible externalities and queue disciplines. *Management Science* **44** 850–858.
- Haviv, Moshe, Binyamin Oz. 2016. Regulating an observable m/m/1 queue. *Operations Research Letters* **44**(2) 196–198.
- Herlitz, J., S. Aune, A. Bng, M. Fredriksson, A. B. Thorn, L. Ekstrm, S. Holmberg. 2005. Very high survival among patients defibrillated at an early stage after in-hospital ventricular fibrillation on wards with and without monitoring facilities. *Resuscitation* **66**(2) 159–66.
- Heyman, D. P., S. Stidham. 1980. The relation between customer and time averages in queues. *Operations Research* **28**(4) 983–994.
- Hoot, Nathan R, Dominik Aronsky. 2008. Systematic review of emergency department crowding: causes, effects, and solutions. *Annals of emergency medicine* **52**(2) 126–136.
- Huang, Junfei, Boaz Carmeli, Avishai Mandelbaum. 2015. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research* **63**(4) 892–908.
- Imbens, G., D.’ Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- Kaiser. 2014. Hospital adjusted expenses per inpatient day by ownership. *Kaiser Family Foundation* <http://kff.org/other/state-indicator/expenses-per-inpatient-day-by-ownership>.
- KC, D., C. Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *M&SOM* **14** 50–65.
- Koole, G. 1997. Assigning a single server to inhomogeneous queues with switching costs. *Theoretical Computer Science* **182** 203–216.

-
- Laskowski, M., R. D. McLeod, M. R. Friesen, B. W. Podaima, A. S. Alfa. 2009. Models of emergency departments for reducing patient waiting times. *PLoS ONE* e6127.
- Lin, D., J. Patrick, F. Labeau. 2014. Estimating the waiting time of multi-priority emergency patients with downstream blocking. *Health care Manag Sci* **71(1)** 88–99.
- McCarthy, M. L., S. L. Zeger, R. Ding, S. R. Levin, J. S. Desmond, J. Lee, D. Aronsky. 2009. Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients. *Annals of Emergency Medicine* **54:4** 492–503.
- Mendelson, Haim, Seungjin Whang. 1990. Optimal incentive-compatible priority pricing for the m/m/1 queue. *Operations research* **38(5)** 870–883.
- Morley, Claire, Maria Unwin, Gregory M Peterson, Jim Stankovich, Leigh Kinsman. 2018. Emergency department crowding: A systematic review of causes, consequences and solutions. *PloS one* **13(8)** e0203316.
- Murray, Michael Bullard, Michael, Eric Grafstein. 2014. Revisions to the canadian emergency department triage and acuity scale implementation guidelines. *Canadian Journal of Emergency Medicine* **6.6** 421–427.
- NQMC. 2016. Emergency department: median time from ed arrival to provider contact for ed patients. National Quality Measures Clearinghouse (NQMC).
- Nørgaard, Birgitte, Christian Mogensen. 2012. Blood sample tube transporting system versus point of care technology in an emergency department; effect on time from collection to reporting? a randomised trial. *Scandinavian journal of trauma, resuscitation and emergency medicine* **20**. doi: 10.1186/1757-7241-20-71.
- Plambeck, Erica L, Qiong Wang. 2013. Implications of hyperbolic discounting for optimal pricing and scheduling of unpleasant services that generate future benefits. *Management Science* **59(8)** 1927–1946.
- Qiu, Y., G. Allon, A. Bassamboo. 2017. How do delay announcements shape customer behavior? an empirical study. *Management Science* **63(1)** 1–20.
- Saghafian, S., W.J. Hopp, M.P. Van Oyen, J.S. Desmond, S.L. Kronick. 2014. Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing and Service Operations Management* **16** 329–345.
- Saghafian, Soroush, Wallace J Hopp, Mark P Van Oyen, Jeffrey S Desmond, Steven L Kronick. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* **60(5)** 1080–1097.
- Schull, M. J., A. Kiss, J. Szalai. 2007. The effect of low-complexity patients on emergency department waiting times. *Annals of Emergency Medicine* **49:3** 257–264.
- Siddharthan, K., W. J. Jones, J. A. Johnson. 1996. A priority queuing model to reduce waiting times in emergency care. *Int J Health Care Qual Assur* 10–16.

- Singh, Amandeep, Kartik Hosanagar, Amit Gandhi. 2020. Machine learning instrument variables for causal inference. *Proceedings of the 21st ACM Conference on Economics and Computation*. 835–836.
- Sobel, Michael E. 2006. What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association* **101**(476) 1398–1407. doi:10.1198/016214506000000636. URL <https://doi.org/10.1198/016214506000000636>.
- Soremekun, Olanrewaju A, Frances S Shofer, David Grasso, Angela M Mills, Jessica Moore, Elizabeth M Datner. 2014. The effect of an emergency department dedicated midtrack area on patient flow. *Academic Emergency Medicine* **21**(4) 434–439.
- Thistlethwaite, D. L., D. T. Campbell. 1960. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology* **51** 309–317.
- Tversky, Amos, Daniel Kahneman. 1991. Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics* **106**(4) 1039–1061.
- Vertesi, Les. 2004. Does the canadian emergency department triage and acuity scale identify non-urgent patients who can be triaged away from the emergency department? *Canadian journal of emergency medicine* **6**(5) 337–342.
- Wolff, Ronald W. 1989. *Stochastic modeling and the theory of queues*. Pearson College Division.
- Xu, Kuang, Carri W. Chan. 2016. Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management* **18**(3) 314–331.
- Yu, Q., G. Allon, A. Bassamboo. 2017a. How do delay announcements shape customer behavior? an empirical study. *Management Science* **63**(1) 1–20.
- Yu, Q., G. Allon, A. Bassamboo. 2017b. The reference effect of delay announcements: A field experiment. *Working paper*.
- Yu, Qiuping, Yiming Zhang, Young-Pin Zhou. 2020. Delay information in virtual queues: A large-scale field experiment on a ride-sharing platform. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3687302.
- Zane, Richard D. 2007. Are low-acuity patients clogging up the ed? *NEJM Journal Watch* Reviewing Schull et al. 2007 Ann Emerg Med 2007 Mar.

Appendix A: Estimation Results for Hospitals 1, 3, 4 and, in 2019 data, SMMC

Figure 5 shows our estimate of the expected truncated externality X_τ for Hospitals 1, 2, 3 and, using the 2019 data set at SMMC. This reinforces the insight from Figure 2 that an additional LAP arrival has a substantial effect on HAP waiting. The effect size is smaller in the SMMC 2019 data than the SMMC 2015 data, consistent with the smaller HAP arrival rate in 2019 than 2016.

For SMMC in 2019, correcting for omitted variable bias (using quasi-randomized wait time information as an instrumental variable) triples the estimated effect size. In partnership with SMMC, we displayed information about the LAP wait time (from registration to start of treatment) on a screen at registration,

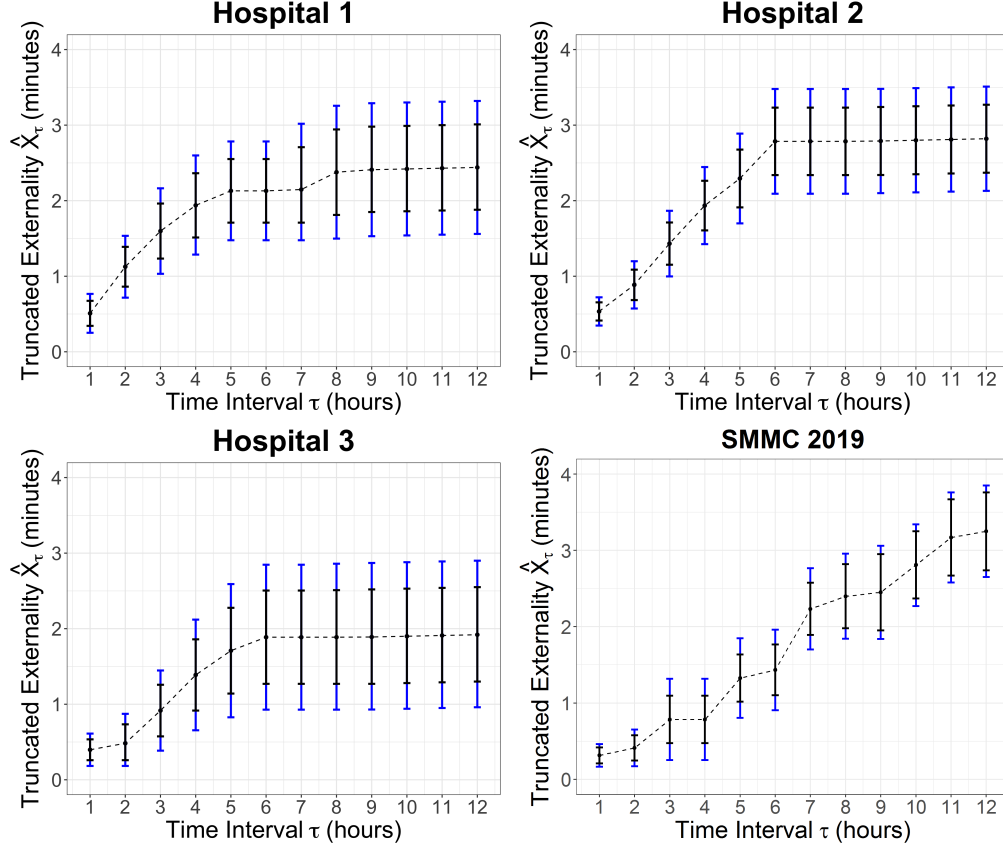


Figure 5 Estimated expected truncated externality \hat{X}_τ (dot) with bootstrap 95% confidence interval (inner error bars) and Bonferroni-corrected 95% confidence interval (outer error bars).

and while doing so collected data for 7,258 patient visits. We displayed the Q-Lasso estimate of the LAP wait time rounded to the nearest multiple of 10 minutes. For example, a Q-Lasso estimate of 16 minutes would be rounded to 20 minutes, and a Q-Lasso estimate of 22 minutes also would be rounded to 20 minutes. With Δ denoting the displayed LAP wait time minus the Q-Lasso estimated LAP wait time, we replicated the IV analysis of §5. The coefficient of Δ in regression model (7) is 0.03 (0.05), and we again leverage the machine learning algorithm of Singh et al. (2020) to strengthen the IV based on Δ . Adopting 2SLS, Figure 6 shows how an additional LAP arrival increases expected post-triage waiting for HAPs that arrive in time interval τ ranging from 1 to 12 hours thereafter. In particular, the expected externality truncated to τ of 12 hours is 9.6 minutes with mean 95% bootstrap confidence interval [8.9, 10.3]. Without using the IV to correct for omitted variable bias, the estimate would be substantially lower, only 3.5 minutes with mean 95% bootstrap confidence interval [2.7, 4.3] in the same limited data set with 7,258 patient visits in 2019, or, as reported above, 3.2 minutes with bootstrap 95% confidence interval [2.7, 3.7] using the full 2019 data set. An important caveat is that inflation Δ in the wait time for LAPs displayed at registration might influence a patient's decision to leave the ED without registering and thereby (because a patient that leaves without registering is not represented in our data set) introduce sample selection bias in the 2019 IV results. Fortunately, our 2015 data does not have this problem because, in 2015, wait time information was provided to patients only after they had registered and started triage. Our main conclusion from IV analysis in the

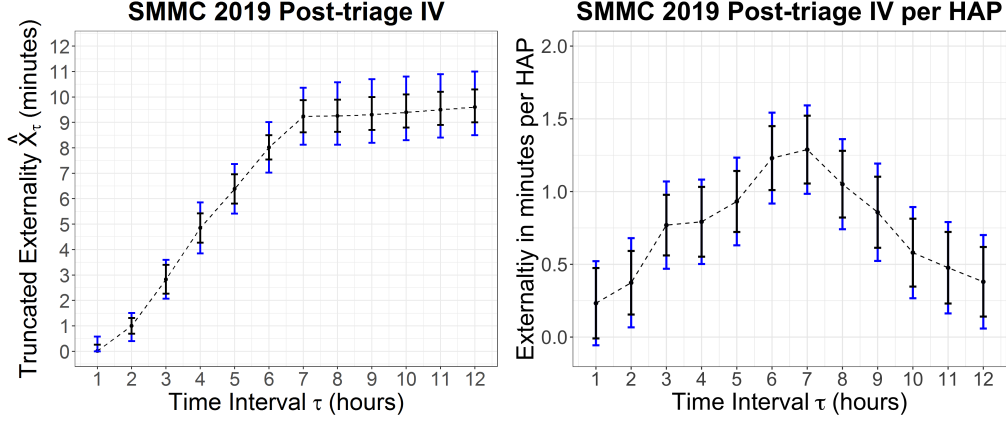


Figure 6 Mean externality (dot); mean 95% bootstrap confidence interval (inner error bars); mean 95% Bonferroni-corrected confidence interval (outer error bars)

2019 data is the same as in the 2015 data: the effect size is substantially larger than estimated based on observational data alone.

Appendix B: Proof of Theorem 1

First, 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 ($i \in \{1, 2, \dots\}$ 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 $j = 1$ so we drop the subscript j when $j = 1$.

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

$$W^{(i)} = V(t^{(i)}) + R_D(t^{(i)}) + D(t^{(i)}). \quad (11)$$

Therefore,

$$w = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n V(t^{(i)}) + \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n R_D(t^{(i)}) + \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n D(t^{(i)}), \quad (12)$$

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,

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n V(t^{(i)}) = \lim_{t \rightarrow \infty} t^{-1} \int_0^t V(u) du \text{ and } \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n R_D(t^{(i)}) = \lim_{t \rightarrow \infty} t^{-1} \int_0^t R_D(u) du, \quad (13)$$

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

$$\lim_{t \rightarrow \infty} t^{-1} \int_0^t V(u) du = \lambda_1 \mathbb{E}[S_1] w + \lambda_1 \mathbb{E}[S_1^2]/2. \quad (14)$$

REMARK 1. Example 5-21 on page 291 of (Wolff 1989) applies that theorem to to obtain (14) 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 ω 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 time $W^{(i)}$ for each i as well as the arrival process, (c) service times $S^{(i)}$ are i.i.d, (d) $\lim_{n \rightarrow \infty} n/t^{(n)} = \lambda_1$ for sample path ω , and (e) as $n \rightarrow \infty$, $n^{-1} \sum_{i=1}^n G_i$ converges to a finite limit for sample path ω , where $G_i = S^{(i)} W^{(i)} + [S^{(i)}]^2/2$. In light of the assumptions we made in §6 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 $R_D(t)$, $\mathbb{I}[R_D(t) > 0]$, and the customer average of $D(t^{(i)})$.

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

- (a) $\lim_{t \rightarrow \infty} t^{-1} \int_0^t R_D(u) du = \lambda_1 \rho_{2p} \mathbb{E}[D^2]/2,$
- (b) $\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n D(t^{(i)}) = \rho_{2p} \mathbb{E}[D],$ and
- (c) $\lim_{t \rightarrow \infty} t^{-1} \int_0^t \mathbb{I}[R_D(u) > 0] du = \lambda_1 \rho_{2p} \mathbb{E}[D].$

Combining (12), (14), and parts (a)-(b) of Lemma 1, we obtain

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

Solving for w gives,

$$w = \frac{\lambda_1 \frac{\mathbb{E}[S_1^2]}{2} + \rho_{2p} \mathbb{E}[D] + \lambda_1 \rho_{2p} \frac{\mathbb{E}[D^2]}{2}}{1 - \lambda_1 \mathbb{E}[S_1]} = \frac{\lambda_1 \frac{\mathbb{E}[S_1]^2 + \sigma[S_1]^2}{2} + \rho_{2p} \mathbb{E}[D] + \lambda_1 \rho_{2p} \frac{\mathbb{E}[D]^2 + \sigma[D]^2}{2}}{1 - \lambda_1 \mathbb{E}[S_1]}.$$

Next, we prove(9). 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 $w_1 =$

$\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n R_{S_2}(t_1^{(i)}) + \lim_{n \rightarrow \infty} n^{-1} V_1(t_1^{(i)})$, if each limit on the right hand side exists. For each in turn, we will show that the limit exist w. p. 1 and derive an expression for it. PASTA implies that w. p. 1,

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n R_{S_2}(t_1^{(i)}) = \lim_{t \rightarrow \infty} t^{-1} \int_0^t R_{S_2}(u) du \quad \text{and} \quad \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n V_1(t_1^{(i)}) = \lim_{t \rightarrow \infty} t^{-1} \int_0^t V_1(u) du,$$

if the time average limits exist. Given that (14) 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,

$$\lim_{t \rightarrow \infty} t^{-1} \int_0^t V_1(u) du = \lambda_1 \mathbb{E}[S_1] w + \lambda_1 \mathbb{E}[S_1^2]/2. \quad (15)$$

For, $\lim_{t \rightarrow \infty} t^{-1} \int_0^t R_{S_2}(u) du$, we can use the following variant of Lemma 1(a).

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

Combining (15) and Lemma 2, we obtain $w_1 = \lambda_1 \mathbb{E}[S_1] w_1 + \lambda_1 \mathbb{E}[S_1^2]/2 + \rho_{2n} \frac{\mathbb{E}[S_2^2]}{2\mathbb{E}[S_2]}$, that gives,

$$w_1 = \frac{\lambda_1 \mathbb{E}[S_1^2]/2 + \rho_{2n} \frac{\mathbb{E}[S_2^2]}{2\mathbb{E}[S_2]}}{1 - \lambda_1 \mathbb{E}[S_1]} = \frac{\lambda_1 (E[S_1]^2 + \sigma[S_1]^2)/2 + \frac{\rho_{2n}}{\mathbb{E}[S_2]} (\mathbb{E}[S_2] + \sigma[S_2]^2)/2}{1 - \lambda_1 \mathbb{E}[S_1]}. \quad \square$$

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 $I_{1 \rightarrow 2} \subset \{1, 2, \dots\}$. Observe that $D(t^{(i)}) > 0$ if and only if $i \in I_{1 \rightarrow 2}$. 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 ρ_{2p} , the long run average fraction of time that a class 2 customer is in service,

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbb{I}[i \in I_{1 \rightarrow 2}]/n = \lim_{t \rightarrow \infty} \int_0^t \mathbb{I}[D(u) > 0]/t du = \rho_{2p} \quad \text{w. p. 1.} \quad (16)$$

For $i = 1, 2, \dots$ define the function

$$f_i(t) = \begin{cases} D(t^{(i)}) - (t - t^{(i)}) & \text{if } t^{(i)} \leq t < t^{(i)} + D(t^{(i)}), \\ 0 & \text{otherwise.} \end{cases}$$

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 the SLLN and (16) to obtain, w. p. 1,

$$\bar{G} \equiv \lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{G_i}{n} = \lim_{n \rightarrow \infty} \left(\frac{\sum_{i=1}^n \mathbb{I}[i \in I_{1 \rightarrow 2}]}{n} \times \frac{\sum_{i=1}^n D(t^{(i)})^2/2}{\sum_{i=1}^n \mathbb{I}[i \in I_{1 \rightarrow 2}]} \right) = \rho_{2p} \mathbb{E}[D^2]/2. \quad (17)$$

Note that, $\sum_{i=1}^n \mathbb{I}[i \in I_{1 \rightarrow 2}]$ 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,

$$\bar{H} \equiv \lim_{t \rightarrow \infty} t^{-1} \int_0^t H(u) du = \lim_{t \rightarrow \infty} t^{-1} \int_0^t R_D(u) du, \quad (18)$$

assuming the limits in (18) exist.

Combining (17), (18), and Theorem 1 of (Heyman and Stidham 1980), we obtain that, w. p. 1, the limit \bar{H} in (18) 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 (17), i.e., w. p. 1,

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{D(t^{(i)})}{n} = \lim_{n \rightarrow \infty} \left(\frac{\sum_{i=1}^n \mathbb{I}[i \in I_{1 \rightarrow 2}]}{n} \times \frac{\sum_{i=1}^n D(t^{(i)})}{\sum_{i=1}^n \mathbb{I}[i \in I_{1 \rightarrow 2}]} \right) = \rho_{2p} \mathbb{E}[D].$$

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, $\bar{G} = \rho_{2p} \mathbb{E}[D]$, and $\bar{H} = \lim_{t \rightarrow \infty} \int_0^t \mathbb{I}[R_D(u) > 0] du / t$. Therefore, the identity $\bar{H} = \lambda_1 \bar{G}$ holds w. p. 1, proving (c). \square

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, $\rho_{2n} \equiv \lim_{t \rightarrow \infty} t^{-1} \int_0^t \mathbb{I}[R_{S_2}(u) > 0] du$ 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 $\lim_{t \rightarrow \infty} N_2(t) = \infty$ 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,

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

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 $\lim_{t \rightarrow \infty} \left[\frac{\sum_{i=1}^{N_2(t)} S_2^{(i)}}{N_2(t)} \right] \frac{N_2(t)}{t}$ and applying SLLN establishes that w.p. 1 $\lim_{t \rightarrow \infty} \sum_{i=1}^{N_2(t)} S_2^{(i)} / N_2(t) = \mathbb{E}[S_2]$ so $\lim_{t \rightarrow \infty} t^{-1} N_2(t)$ 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,

$$\bar{G} \equiv \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n G_i = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n (S_2^{(i)})^2 / 2 = \mathbb{E}[S_2^2] / 2. \quad (19)$$

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,

$$\bar{H} \equiv \lim_{t \rightarrow \infty} t^{-1} \int_0^t H(u) du = \lim_{t \rightarrow \infty} t^{-1} \int_0^t R_{S_2}(u) du. \quad (20)$$

Finally, we note that since $S_2^{(i)}$ are i.i.d. with finite expectation the “technical assumption” required to use Theorem 1 of Heyman and Stidham (1980) is also satisfied. Combining that result with (19) and (20), we have $\bar{H} = (\rho_{2n} / \mathbb{E}[S_2]) \bar{G}$ w. p. 1. That finishes the proof. \square

Electronic Companion for “LAPs Delay HAPs in an Emergency Department”

Appendix EC.1: Per HAP Externality Estimation for Hospital 1, 2, 3, and SMMC in 2019

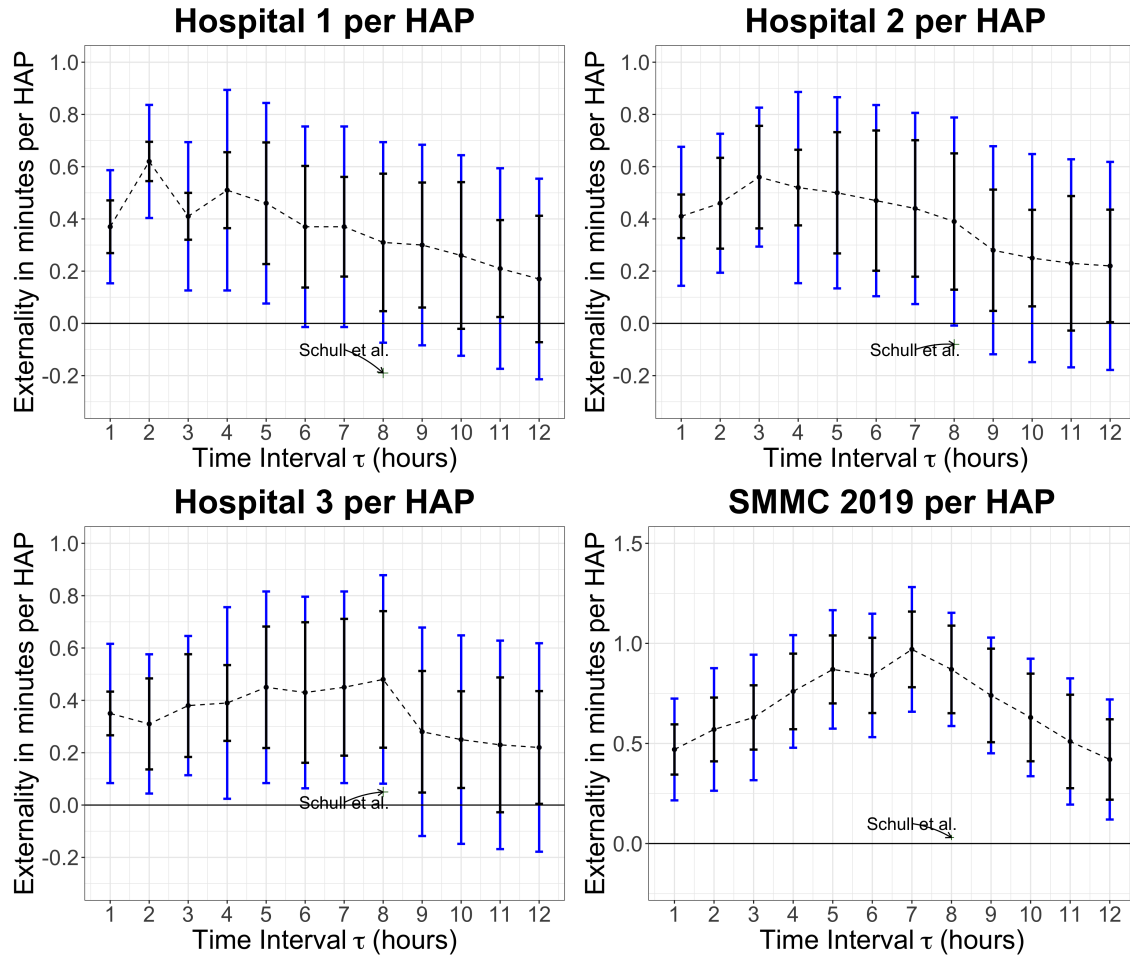


Figure EC.1 Estimated expected truncated externality \hat{X}_τ (dot) with bootstrap 95% confidence interval (inner error bars) and Bonferroni-corrected 95% confidence interval (outer error bars).

Appendix EC.2: Robustness Check for the Bootstrap Confidence Intervals

We performed a robustness check for the bootstrapped confidence interval generation method explained in §4.2. We created 50 independently generated data sets via the simulation set up (with non-stationary arrivals) introduced in §7. Specifically, we created 50 simulation data and for each of them, we draw a time point t from every eight hour interval. This led to 50 independent data

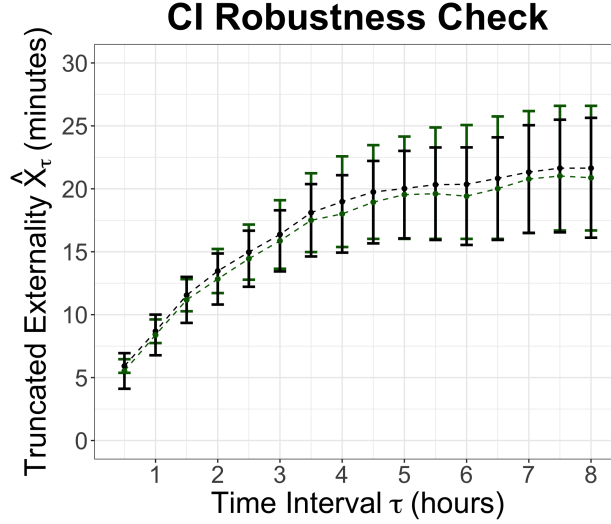


Figure EC.2 Estimated average effect from the bootstrapped approach (black dashed line with solid round dots); 95% bootstrapped confidence interval (error bars in solid black lines); Average effect from the 50 dataset (green dashed line with solid round dots); mean 95% true confidence interval (error bars in solid green lines);

sets to be used as inputs to our estimator. We refer to these 50 data sets as D_1, \dots, D_{50} . Then, we calculated two different sets of confidence intervals:

- *True confidence intervals:* We applied our estimator to each one of D_1, \dots, D_{50} independently and obtained the corresponding estimates for X_τ . Then we calculated a 95% confidence interval from the 50 estimated values. This is the true 95% confidence interval since the 50 data sets are by design independent and identically distributed.
- *Bootstrapped confidence intervals:* We applied our bootstrapped confidence interval method from §4.2 to D_1 to obtain 95% confidence intervals for the X_τ .

Figure EC.2 demonstrates both of these confidence intervals for different values of τ . The difference between the bootstrapped confidence intervals (shaded in black) vs. the true confidence intervals (shaded in green) is very small. While this is not a formal statistical test, but it is a rigorous visual inspection that the bootstrapped confidence intervals are valid.

Appendix EC.3: Explanatory Variables \vec{C} in each Hospitals

The complete list of explanatory variables for hospitals includes

- Time related variables includes the period of day (pod) indicators: pod24, pod12, pod8, pod6, pod4, pod3, pod2, the numerical number following “pod” indicates the number of sub-intervals a day breaks into; e.g., pod24 is a categorical variables with 24 levels indicating whether a patient’s arrival time (equivalent to the registration time) falls either during 00:00 - 01:00, 01:00 - 02:00, ..., or 23:00 - 00:00. Another variable is a binary variable for Fast Track indicating whether a patient’s arrival time falls into the time period when Fast Track is on. In SMMC,

the Fast Track is in effect during the weekdays between 12 : 00 noon - 10 : 00 PM, and in Hospital 1, 2, and 3, the Fast Track is in effect during the weekdays between 8 : 00 AM - 11 : 00 PM. Other variables are month, season and weekend indicators, which indicate whether a patient's arrival time falls into the corresponding month, season (4 levels, January - March, April - June, July - September, or October - December), and whether the patient arrives during weekend.

- Average number of nurses and physicians at the time of a patient's arrival.
- Number of LAPs and HAPs currently in the ED (the sum of those that are currently in queue and those that are currently in service) at the time of a patient's arrival.
- Indicator variables for weather information: is.rainy, is.foggy, and temperature at the moment of patient's arrival.

The goodness of fit of the model on majority of the bootstrapped datasets is of $R^2 \approx 0.85$. Table EC.2 shows the final set of selected controls for each hospital. We provide summary of statistics the these controls for the SMMC 2015 experiment data in Table EC.1. In each bootstrapped dataset, we have a total of 142 observations.

Table EC.1 Summary of statistics for selected variables for the regression data using SMMC 2015 experiment with $\epsilon = 10$ minutes and $\tau = 8$ hours.

Variable	Definition	Mean (Standard Deviation)
η_τ	sum of HAPs wait time in τ	48.6 (72.4)
N_ϵ^L	# of LAP arrivals in ϵ interval	0.74 (1.2)
N_τ^L	# of LAP arrivals in τ interval	23.8 (19.4)
N_τ^H	# of HAP arrivals in τ interval	2.1 (2.5)
$N_\epsilon^L \cdot N_\tau^H$	cross term	2.8 (5.2)
$N_\epsilon^L \cdot N_\tau^L$	cross term	29.4 (48.3)
$N_\tau^L \cdot N_\tau^H$	cross term	87.5 (127.3)
$N_\epsilon^L \cdot N_\tau^L \cdot N_\tau^H$	cross term	119.4 (244.4)
pod3	period of the day	0: 45 (6) 1: 49(5) 2: 48(5)
S_t^H	# of HAPs in the system at time t	2.3 (2.1)
S_t^L	# of HAPs in the system at time t	15.6 (9.7)
is.weekend	indicating if the observation falls in the weekend	True: 36 , False: 106
Inflation	the average inflation as defined in §5	10.3 (7.7)

Appendix EC.4: Robustness Check for the Approach Estimating the Effect of Fast Track in §4.3

In San Mateo Medical Center (SMMC), we know that the Fast Track is in effect from 12 - 10 pm. For the Table EC.3 Fast Track results, we randomly select t from 12-10 pm and excluding observations if LAPs arrived to triage between 11:50-12:00 but started treatment before 12:00.

Table EC.2 The selected explanatory variables \vec{C} for each hospitals: all selected explanatory variables are time-related categorical variables

Hospitals	\vec{C}
SMMC	pod3, S_t^H , S_t^L , weekend
Hospital 1	pod3, S_t^H , weekend, season, Fast Track
Hospital 2	pod3, S_t^H , S_t^L , weekend, season
Hospital 3	pod3, S_t^H , weekend, season
Hospital 4	pod4, S_t^H , S_t^L , weekend, season

We also exclude LAPs that arrived to triage before 10 pm but started treatment after 10 pm . Whereas for the no Fast Track results, we randomly select t from the remaining time interval of 10 pm to 12 pm and excluding observations if LAPs arrived to triage between 9:50 - 10:00 pm but started treatment before 10 pm and also excluding LAPs arrived to triage before 12 pm but started treatment after 12 pm.

Table EC.3 Estimated effect of additional LAP that arrives during 12-10 pm (Fast Track) vs. 10-12 pm (No Fast Track) on HAPs' expected pre- and post-triage wait time \hat{X}_8 at SMMC.

Pre-triage No FT	Pre-triage FT	Post-triage No FT	Post-triage FT
4.3(± 1.8)	4.4(± 1.5)	5.6(± 1.8)	5.1(± 1.7)

Appendix EC.5: Replicating Schull et al. (2007)'s analysis using SMMC, Hospital 1, 2, 3, and 4 Data

In this section, we replicate Schull et al.'s analysis using SMMC, Hospital 1, 2, 3 and 4 data, and present the results in Table EC.4. We redefine LAP, exactly as how Schull et al. do, their notion of low-complexity patients, which is according to 3 factors: a Canadian Triage Acuity Score of less urgent (level 4) or non-urgent (level 5) (Murray and Grafstein 2014), ED arrival not by ambulance, and discharged to home. SMMC data has information on ways of patient's arrival and disposition; however hospital 1, 2, 3, and 4 do have them. Thus in hospital 1, 2, 3 and 4, we label ESI 4 and 5 patients as LAPs.

Table EC.4 The estimated effect of a low-complexity patient (with ESI 4 and 5) arriving per 8 hour interval on the total wait time to treatment (**Total**) for medium- and high- complexity patients (with ESI 1, 2 and 3), and on their mean wait time to treatment (**Per HAP**), using Schull et al.’s method, and our method; the (**Total**) effect focusing on the portion of the estimated effect on the pre-triage delay vs. the post-triage delay when the Fast Track is implemented or not.

		Estimated Externality Caused by a LAP in 8 Hour					
		SMMC 2015	SMMC 2019	Hospital 1	Hospital 2	Hospital 3	Hospital 4
Total	Ours	29.4(±5.9)	6.2(±2.1)	66.2(±4.6)	24.6(±2.2)	5.4(±0.4)	10.2(±1.1)
	Pre-triage No FT	23.7 (±6.7)	4.9 (±1.7)	no info	no info	no info	no info
	Pre-triage FT	19.7 (±4.2)	3.4 (±1.2)	no info	no info	no info	no info
	Post-triage No FT	5.6 (±1.2)	1.2 (±0.5)	no info	no info	no info	no info
	Post-triage FT	4.1 (±1.2)	1.1 (±0.5)	no info	no info	no info	no info
	Schull	-1.8	0.7	-35.1	-0.5	1.1	0.2
Per HAP	Ours	1.5(±0.5)	0.7(±0.2)	0.6(±0.2)	1.5(±0.1)	0.7(±0.2)	1.4(±0.3)
	Schull	-0.10(±0.02)	0.07(±0.02)	-0.67(±0.06)	-0.01(±0.03)	0.03(±0.02)	0.01(±0.02)

Appendix EC.6: Supplements for the Experiment Results from §5

In this section we present supporting material for the pseudo-randomized experiment from §5. In §EC.6.1 we provide empirical evidence for the mechanism behind the effect of inflation Δ on number of low-acuity patients N_ϵ^L . Details of the MLIV procedure for strengthening the IV and obtaining $f_*(\Delta)$ is presented in §EC.6.2 and complete output of the 2SLS using this stronger IV is shown in §EC.6.3.

Here we supply a discussion on the percentage of LAPs might see a different displayed wait time in triage. We define patients’ triage time to be the time interval between triage time-stamps. The mean of patients’ triage time in SMMC is 8 minutes. Our displayed wait time is refreshed every 10 minutes. We investigate whether patients experience a wait time update while in triage and whether the displayed wait time is different. We found that 70% of LAPs indeed experience a wait time update while in triage; However, among them, only about 7% see a different displayed wait time, which means that 95% of all LPAs do not see a change in displayed wait time while in triage.

Moreover, the above 95% might be an underestimation. The mean of patient’s triage time may be less than 8 minutes since the time interval between triage time-staps does not exclude time spent on patients moving from the triage room to the waiting room and back for the next patient (changeover time). If we exclude this 1-2 minutes of changeover time, a shorter triage time means that fewer patients would experience the wait time update.

EC.6.1. Intuition on How Δ Impacts N_ϵ^L

Here we provide empirical support for the hypothesis (discussed in §5) that a larger inflation Δ in the displayed wait time increases the number of patients who wait for long enough to start

treatment. The logistic regression (EC.1), which controls for triage-level indicators and a set of time-related indicators \vec{T} , yields an inflation effect of $\beta_1 = -0.07$ (0.03)* on probability that a patient leaves without being seen (LWBS).

$$P(\text{LWBS}) = \beta_0 + \beta_1 \Delta + \beta_2 \mathbb{I}\{\text{ESI} = 4\} + \beta_3 \mathbb{I}\{\text{ESI} = 5\} + \vec{\beta}_4 \cdot \vec{T} \quad (\text{EC.1})$$

We fit (EC.1) on data from the 5265 LAPs who arrived while the triage room screen was fully operational. The treatment variable Δ denotes the inflation and \vec{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 LAP leaving without being seen. Table EC.5 reports estimates of coefficients for Δ obtained via logit, probit and linear regression, as well as the corresponding estimates obtained when replicating the analysis on low-acuity ED visits for which the screen was off.

EC.6.2. Details in Adopting the “MLIV” Method

Following the recipe in (Singh et al. 2020), we perform the steps below to strengthen our instrument to correct our estimate for the externality on each bootstrap dataset \mathcal{S} .

1. **Outer Loop:** split the data \mathcal{S} (roughly) equally into a 2-fold partition, such that each partition \mathcal{S}_k ($k \in \{1, 2\}$) has size $\lfloor \frac{n}{2} \rfloor$ or $\lfloor \frac{n}{2} \rfloor + 1$. For each partition k , define \mathcal{S}_k^c to be the excluded data.
2. **Inner Loop:** for each partition k , we learn the optimal instrument function $f_k(\cdot; \zeta_k)$ to predict N_L^c using the excluded data \mathcal{S}_k^c . Specifically, $f(\Delta; \zeta)$ is a class of instrumental variable functions parameterized by ζ , here Δ is the “Inflation”. The pool of candidate f functions includes XGBoost², Lasso, Ridge, and Elastic Net, and the hyperparameters pertinent to f is tuned through 2-fold cross-validation.
3. Compare prediction accuracy (using RMSE metric) of each of the two instrument functions $f_k(\Delta; \zeta_k)$ and choose the one that is more accurate. Denote the selected function and hyperparameter by $f_*(\Delta; \zeta_*)$.
4. Now, for the entire dataset \mathcal{S} , we can obtain the new instrument $f_*(\Delta, \zeta_*)$ that we also denote by $f_*(\Delta)$ for simplicity. Thus we can use the standard 2SLS method to get the estimation for the externality.

Using 2015 experiment data, The left sub-figures in Figures EC.4 and EC.3 show the boxplots (the median, the 25th percentile, and the 75th percentile) of the p-values of the IV using weak-instrument test and the Wu-Hausman test, respectively, across all 20 bootstrapped datasets at $\tau = 8$ hours, before and after the implementation of “MLIV”. The results demonstrate that “MLIV” method strengthens the IV. Similar boxplots for 2019 experiment (as discussed in §5) are shown in the right sub-figures of Figures EC.4 and EC.3.

²XGBoost and Elastic Net method performs better in minimizing the RMSE in majority of bootstrapped datasets, and there no much different between the two. Thus we use XGBoost as our main machine learning method.

	Experiment			Control (Screen Off)		
	<i>Logit</i>	<i>Probit</i>	<i>OLS</i>	<i>Logit</i>	<i>Probit</i>	<i>OLS</i>
(Intercept)	-4.06*** (1.07)	-2.13*** (0.44)	0.02 (0.018)	-22.35 (3830.85)	-6.48 (526.9)	-0.007 (0.017)
Forecast Inflation	-0.07* (0.03)	-0.03* (0.01)	-0.001* (0.001)	0 (0.08)	-0.01 (0.03)	0 (0.001)
Triage Level = 4	1.1** (0.34)	0.46*** (0.14)	0.016** (0.005)	1.38*** (0.41)	0.55*** (0.16)	0.017*** (0.005)
Triage Level = 5	1.9*** (0.52)	0.8*** (0.24)	0.042*** (0.012)	1.67* (0.7)	0.69* (0.3)	0.022 (0.012)
Arrived during 1am-2am	0.07 (1.44)	0.03 (0.6)	0.002 (0.026)	17.95 (3830.85)	4.25 (526.9)	0.026 (0.025)
Arrived during 2am-3am	0.48 (1.44)	0.15 (0.62)	0.009 (0.027)	0.09 (6100.27)	0.05 (840.03)	0.001 (0.025)
Arrived during 3am-4am	-16.42 (3615.04)	-3.86 (509.07)	-0.016 (0.032)	0.35 (6335.17)	0.14 (872.19)	0.004 (0.026)
Arrived during 4am-5am	-16.52 (3220.8)	-3.91 (453.59)	-0.018 (0.03)	-0.06 (7443.16)	-0.01 (1025.46)	-0.001 (0.031)
Arrived during 5am-6am	0.71 (1.44)	0.35 (0.6)	0.016 (0.029)	0.18 (6389.93)	0.07 (878.88)	0.002 (0.026)
Arrived during 6am-7am	-16.3 (2999.83)	-3.85 (420.72)	-0.014 (0.028)	-0.05 (5331.74)	0 (734.02)	-0.001 (0.022)
Arrived during 7am-8am	-16.7 (1973.44)	-3.99 (277.74)	-0.02 (0.023)	-0.23 (5005.23)	-0.05 (690.04)	-0.003 (0.021)
Arrived during 8am-9am	-16.87 (1642.71)	-4.07 (230.84)	-0.024 (0.021)	-0.17 (4536.66)	-0.04 (624.66)	-0.002 (0.019)
Arrived during 9am-10am	-1.44 (1.43)	-0.57 (0.56)	-0.016 (0.02)	16.96 (3830.85)	3.86 (526.9)	0.01 (0.018)
Arrived during 10am-11am	-0.77 (1.24)	-0.33 (0.5)	-0.01 (0.02)	17.77 (3830.85)	4.2 (526.9)	0.024 (0.018)
Arrived during 11am-12pm	-1.7 (1.43)	-0.62 (0.54)	-0.018 (0.019)	17.01 (3830.85)	3.87 (526.9)	0.01 (0.018)
Arrived during 12pm-1pm	0.31 (1.1)	0.14 (0.46)	0.011 (0.02)	16.49 (3830.85)	3.74 (526.9)	0.006 (0.019)
Arrived during 1pm-2pm	0.09 (1.14)	0.02 (0.47)	0.003 (0.02)	17.44 (3830.85)	4.06 (526.9)	0.017 (0.018)
Arrived during 2pm-3pm	0.43 (1.1)	0.18 (0.46)	0.013 (0.02)	-0.06 (4466)	0 (612.34)	-0.001 (0.018)
Arrived during 3pm-4pm	0.52 (1.09)	0.26 (0.45)	0.016 (0.02)	18.27 (3830.85)	4.42 (526.9)	0.039* (0.019)
Arrived during 4pm-5pm	-16.77 (1369.86)	-4 (192.98)	-0.022 (0.02)	0.02 (4477.88)	0.03 (616.27)	0 (0.018)
Arrived during 5pm-6pm	0.7 (1.1)	0.32 (0.46)	0.021 (0.02)	17.89 (3830.85)	4.24 (526.9)	0.027 (0.019)
Arrived during 6pm-7pm	-1.18 (1.43)	-0.45 (0.56)	-0.013 (0.02)	-0.09 (4491.04)	-0.02 (618.15)	-0.001 (0.019)
Arrived during 7pm-8pm	-0.15 (1.17)	-0.08 (0.48)	-0.002 (0.02)	17.86 (3830.85)	4.23 (526.9)	0.025 (0.018)
Arrived during 8pm-9pm	-0.15 (1.17)	-0.04 (0.48)	-0.002 (0.02)	17.86 (3830.85)	4.24 (526.9)	0.025 (0.019)
Arrived during 9pm-10pm	-0.36 (1.24)	-0.1 (0.5)	-0.005 (0.02)	18.32 (3830.85)	4.4 (526.9)	0.035 (0.019)
Arrived during 10pm-11pm	0.5 (1.17)	0.26 (0.49)	0.012 (0.022)	16.91 (3830.85)	3.81 (526.9)	0.01 (0.02)
Arrived during 11pm-12pm	-16.66 (1927.64)	-4 (270.64)	-0.02 (0.023)	17.31 (3830.85)	3.99 (526.9)	0.015 (0.021)

Table EC.5 Coefficient estimates for regression (EC.1).

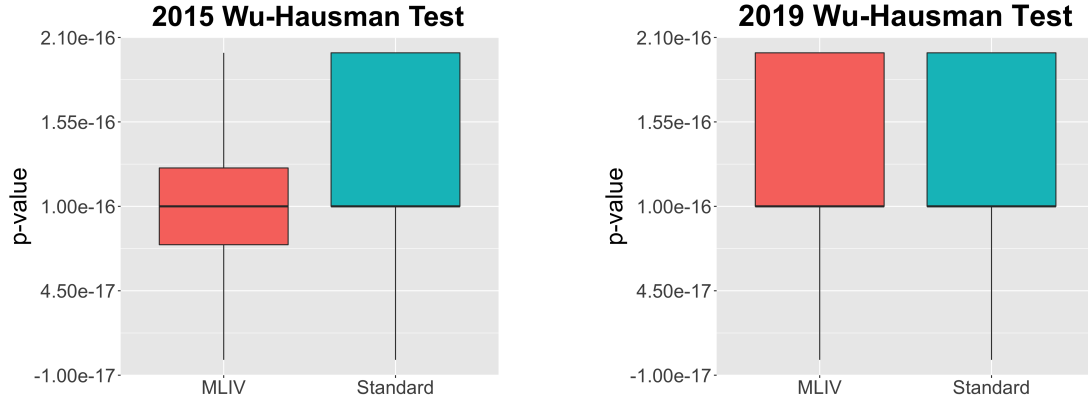


Figure EC.3 Left: boxplots for the p-values of the Wu-Hausman test for the IV using the standard 2SLS method vs. the MLIV method, across bootstrapped dataset, using 2015 experiment data. Right: same analysis as in the Left, but using 2019 experiment data.

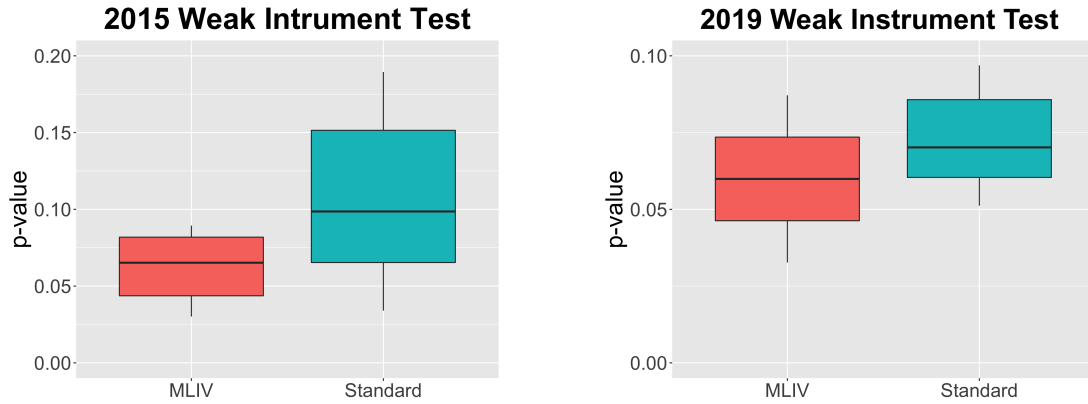


Figure EC.4 Left: boxplots for the p-values of the weak instrument test for the IV using the standard 2SLS method vs. the MLIV method, across bootstrapped dataset, using 2015 experiment data. Right: same analysis as in the Left, but using 2019 experiment data.

EC.6.3. Output of the 2SLS

Table EC.6 presents coefficients of the 2SLS using $f_*(\Delta)$ as IV. Definition of the controls can be found in §4 and §EC.3. The standard error for each estimated coefficient is based on estimates from 20 bootstrapped datasets using SMMC 2015 data. We use \hat{N}_ϵ^L to denote the estimated N_ϵ^L using $f_*(\Delta)$ as IV in the first stage (with same controls as in model (7)) of the 2SLS.

Table EC.6 Estimation Results for 2SLS for regression (2)

Variable	Estimated coefficient (SE)
\hat{N}_ϵ^L	2.2 (0.5)
N_τ^L	0.008 (0.006)
N_τ^H	0.41 (0.13)
$N_\tau^L \times N_\tau^H$	0.001 (0.002)
$\hat{N}_\epsilon^L \times N_\tau^L$	0.02 (0.02)
$\hat{N}_\epsilon^L \times N_\tau^H$	0.43 (0.2)
$\hat{N}_\epsilon^L \times N_\tau^L \times N_\tau^H$	0.006 (0.002)
S_t^H	0.21 (0.09)
S_t^L	0.10 (0.11)
pod31	-0.14 (0.13)
pod32	0.28 (0.11)
is.weekend	0.03 (0.01)
R^2	0.76

Appendix EC.7: Additional Theoretical Results and Robustness Check for our estimator in §7

The triage server itself is a M/M/1 queueing system with arrival rate $\lambda_T = \lambda_H + \lambda_L$, and service rate μ_T . Thus the HAP expected wait time in the triage server, $\mathbb{E}[W_H^T]$, is $\rho_T/(\mu_T - \lambda_T)$, where $\rho_T = \lambda_T/\mu_T$ (Wolff 1989). Moreover, according to Burke's theorem (Burke 1956), we know that the departure process of the triage server is the same as the arrival process. Thus the arrival processes of HAPs and LAPs to the treatment server are still Poisson with arrival rates λ_H and λ_L , respectively. Thus, we can directly get the closed-form expression for the expected HAP wait time in the treatment server, $\mathbb{E}[W_H^M]$, using expression (8) in Theorem 1, which is derived in a more generalized setting. Thus we obtain the closed-form expression for the expected HAP wait time from arrival to treatment, $\mathbb{E}[W_H^T] + \mathbb{E}[W_H^M]$. Now, to get the theoretical value for the estimate in Schull et al. (2007), under each set of parameters, we can calculate a numeric difference in the expected wait time given by

$$\left\{ \mathbb{E} \left[W_H^T \left(\lambda_L + \frac{1}{8 \times 60} \right) \right] + \mathbb{E} \left[W_H^M \left(\lambda_L + \frac{1}{8 \times 60} \right) \right] \right\} - \left\{ \mathbb{E} \left[W_H^T (\lambda_L) \right] + \mathbb{E} \left[W_H^M (\lambda_L) \right] \right\}. \quad (\text{EC.2})$$

Table EC.7 shows the externality estimation with exponential transition delay time with mean 5 minutes to show the results stay qualitatively the same as reported in Table 6.

To evaluate how our estimator performs with realistic diurnal variation in arrival rates for LAP and HAP patients, we modify the arrival rates for LAPs and HAPs in the simulation model. Specifically, within each of the 24 hours in a day, we set λ_L and λ_H to the mean number of

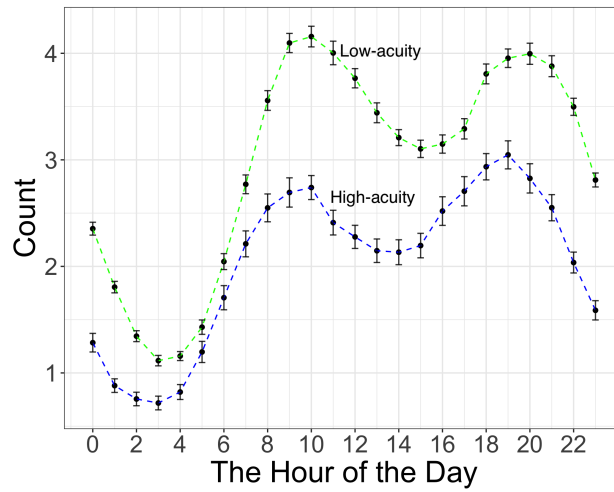


Figure EC.5 Average count of low-acuity arrivals (on the top) and average count of high-acuity arrivals (on the bottom) vs. arrival period in a day; and the error bar centered around each dot is obtained using mean 95% bootstrapped confidence interval.

Table EC.7 Estimated effect (and 95% confidence interval) of one additional LAP arrival in an 8 hour interval on expected wait time *per HAP* and *total* for all HAPs arriving in the 8 hour interval, for the Oracle, our approach and Schull et al.'s approach. Parameters are specified in Table 5 except the transition delay that follows a exponential distribution with a mean of 5 minutes.

		Estimated Externality Caused by a LAP in 8 Hour				
		H1/SMMC	H2	H3	H4	SMMC peak
<i>total</i>	Ours	6.5(± 0.7)	3.5(± 0.6)	2.5(± 0.5)	4.8(± 0.6)	36.5(± 1.8)
	Oracle	5.3(± 1.5)	2.5(± 1.1)	2.2(± 0.6)	6.9(± 2.1)	46.3(± 12.2)
<i>per HAP</i>	Ours	1.1(± 0.4)	0.3(± 0.2)	0.5(± 0.1)	1.4(± 0.1)	8.3(± 0.2)
	Schull	0.34(± 0.01)	0.24(± 0.01)	0.18(± 0.03)	0.38(± 0.02)	0.36(± 0.02)
	Oracle	1.5(± 0.6)	0.3(± 0.2)	0.4(± 0.1)	1.5(± 0.3)	9.3(± 2.2)

LAPs and of HAPs, respectively, that arrive in that hour of the day, in SMMC observational data (demonstrated in Figure EC.5). We replicate the analysis described in §7 with that diurnal variation in arrival rates as well as exponential distributed transition delay time, and summarize the results in Table EC.8. Our approach reasonably estimates the true externality (Oracle) in the non-stationary system whereas Schull et al.'s approach yields an underestimate.

Another robustness check that we did for our estimator is to allow LAPs to renege before the treatment server. Across SMMC, hospital 1, 2, 3, and 4, the left without being seen rate for LAPs ranges from 1% – 3%; among them 99.5% leave after they are triaged. To model this phenomenon, on top of the non-stationary queue setup discussed for Table EC.8, we let LAPs to renege the

Table EC.8 Mean true externality truncated to 8 hours (Oracle), our estimate of the mean externality truncated to 8 hours, and our estimate without using interaction terms (3) in the regression; estimate of the mean externality truncated to 8 hours by extrapolating the per HAP estimate from Schull et al.

Mean Service Time (min.)			Transition Delay	Externality (min.)			
Triage	Treatment HAP	Treatment LAP	Distribution	Oracle	Ours	Ours without (3)	Schull et al.'s
8	18	6	exponential(5)	30.2 (± 12.5)	26.4	35.2	5.3
8	18	6	constant(5)	25.6 (± 10.1)	21.3	31.8	5.0

Table EC.9 The true (Oracle) and the estimated externality of a LAP on HAPs wait time truncated to 8 hours via our estimator for simulation data generated based on the actual high- and low-acuity arrival patterns in SMMC, which is demonstrated in Figure EC.5, further assuming LAPs renegeing with 3% probability after being triaged but before being treated (leave without being seen). Externality is in *minute*, and μ_T , μ_H , and μ_L are in $1/\text{minute}$. The transition delay time is set to be an constant at 5 *minutes*.

	(1) Parameters of the Queue				(2) externality	
	μ_T	μ_H	μ_L	LWBS	Oracle	Ours
SMMC	$\frac{1}{8}$	$\frac{1}{18}$	$\frac{1}{6}$	3%	25.3 (± 9.9)	21.2

system with a probability of 3% after being triaged but before entering the treatment server. We replicate the analysis done for Table EC.8 and summarize the result in Table EC.9.

Appendix EC.8: Numerical Examples Comparing Average Wait Time for HAPs, Under the Non-preemption Policy Versus Under the Preemption Policy.

In row 1, 2, and 4 of Table EC.10, we provide numerical examples demonstrating the average HAPs wait time, under the non-preemption policy, to be smaller than under the preemption policy.

We simulate a M/M/1 queueing system in §6 with the same input parameters as described in Table 5, mimicking SMMC, hospital 1, 2, 3 and 4. We assume the transition delay time $T = 3 + 27 \cdot G$, where G is a random variable following a beta distribution with shape parameters $\alpha = 1$ and $\beta = 3$, so that the time falls in the range $[3, 30]$ minutes, mean 9.75, and variance 27.5.

Table EC.10 The long-term average HAPs wait time with its 95% confidence interval under the preemption policy (\bar{W}_H^P) and under the non-preemption policy (\bar{W}_H^{NP}) using simulation data generated based on actual average ED utilization of SMMC, hospital 1, 2, 3, and 4. The units of the queueing parameters and the wait times are consistent: \bar{W}_H^P and \bar{W}_H^{NP} are in minutes; λ_H , λ_L , μ_H , and μ_L are in minute⁻¹.

	λ_H	λ_L	μ_H	μ_L	\bar{W}_H^P	\bar{W}_H^{NP}
H1/SMMC	$\frac{1}{60}$	$\frac{1}{15}$	$\frac{1}{18}$	$\frac{1}{6}$	14.38(± 0.37)	11.16(± 0.34)
H2	$\frac{1}{55}$	$\frac{1}{30}$	$\frac{1}{18}$	$\frac{1}{6}$	12.15(± 0.33)	10.38(± 0.31)
H3	$\frac{1}{60}$	$\frac{1}{30}$	$\frac{1}{18}$	$\frac{1}{6}$	11.09(± 0.28)	9.46(± 0.26)
H4	$\frac{1}{55}$	$\frac{1}{15}$	$\frac{1}{18}$	$\frac{1}{6}$	15.67(± 0.37)	12.17(± 0.34)