

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

Accurate ED Wait Time Prediction

Erjie Ang*, Sara Kwasnick*, Mohsen Bayati, Erica L. Plambeck
Graduate School of Business, Stanford University, Stanford, California 94305

Michael Aratow

San Mateo Medical Center, San Mateo, California 94403

This paper proposes the Q-Lasso method for wait time prediction, which combines statistical learning with fluid model estimators. In historical data from four remarkably different hospitals, Q-Lasso predicts the emergency department (ED) wait time for low-acuity patients with greater accuracy than rolling average methods (currently used by hospitals), fluid model estimators (from the service OM literature), and quantile regression methods (from the emergency medicine literature). Q-Lasso achieves greater accuracy largely by correcting errors of underestimation in which a patient waits for longer than predicted. Implemented on the external website and in the triage room of the San Mateo Medical Center (SMMC), Q-Lasso achieves over 30% lower mean squared prediction error than would occur with the best rolling average method. The paper describes challenges and insights from the implementation at SMMC.

Key words: OM Practice, Service Operations, Empirical Research, Health Care Management

1. Introduction

A growing number of hospitals are publishing estimates of the Emergency Department (ED) wait time on websites, mobile device applications, billboards and screens within the hospital. The American College of Emergency Physicians (ACEP 2012) has warned that these estimates are misleading to patients, and has called for improved accuracy.

In response, this paper proposes a more accurate, widely-applicable method for predicting the ED wait time, called “Q-Lasso.” In historical data from four different hospitals, the paper evaluates the accuracy of Q-Lasso and plausible alternative methods. Relative to a rolling average, the method currently used by hospitals (Dong et al. 2015), Q-Lasso reduces the

*Student first authors

out-of-sample mean squared error (MSE) in prediction by over 21% on average for the four hospitals. In implementation at the San Mateo Medical Center, Q-Lasso reduces the MSE by over 30%. Nevertheless, the title of the paper is aspirational: even Q-Lasso, optimized for accuracy, makes some large errors.

Q-Lasso combines queuing theory with the Lasso method of statistical learning. Lasso predicts a dependent variable (e.g. wait time) as a linear function of candidate predictor variables (e.g. variables describing the current state of the ED) with the objective of minimizing the mean squared error in the prediction plus a penalty function. More specifically, to avoid overfitting, Lasso penalizes assignment of a nonzero coefficient to a predictor variable, and thus chooses to use only a subset of the candidate predictor variables.

The primary challenge for successful application of Lasso is to initially input a set of candidate predictor variables that includes effective predictor variables (Domingos 2012). On the other hand, the primary challenge for wait time prediction via queueing theory is to effectively model a complex, idiosyncratic ED (Armony et al. 2012, Shi et al. 2012). To surmount both challenges, Q-Lasso incorporates a wide variety of candidate predictor variables that may be interpreted as candidate fluid model estimators, including ones proposed by queueing theorists in the service operations management literature. Intuitively, in assigning coefficients to these fluid-model-inspired candidate predictor variables, Q-Lasso automatically decides how best to estimate the workload that the ED must process before an arriving patient will start treatment, and how best to estimate the corresponding processing rate. Simultaneously, Q-Lasso automatically decides how best to predict the wait time as a weighted combination of fluid model estimators, a conventional rolling average estimator, deterministic estimators reflecting the diurnal variation in wait time, etc. Admittedly, inference about Lasso’s coefficient estimates is poorly understood, despite being an active area of research (Belloni et al. 2014, Javanmard and Montanari 2014); one cannot reliably interpret the coefficients to gain an understanding of how the underlying system operates, unlike in a queueing structural approach exemplified by Armony et al. (2012) and Shi et al. (2012).

Although this paper focuses on EDs, Q-Lasso is potentially applicable to other complex, data-rich service systems.

The Q-Lasso method is motivated by the service operations management literature that proposes accurate and practical methods for wait time prediction. Armony et al. (2009) propose using the average wait time of the last k customers to enter service or, for systems

with diurnal fluctuation in the arrival rate, the historical average wait time for the period of the day. Ibrahim et al. (2014) prove that the former with $k = 1$ is accurate under conditions that commonly arise in call centers. However, Ibrahim and Whitt (2009a) show that a fluid model estimator – that predicts wait time as the ratio of queue length to processing rate – is more accurate than various such estimators based on the recent history of customer wait times, in a $GI/M/s$ queue. Ibrahim and Whitt (2009b, 2011a,b) extend that result, and the fluid model estimator, for more complex systems with abandonment and time-varying arrival and processing rates. These papers motivate the inclusion of deterministic diurnal, rolling average and fluid model estimators (including one in which the processing rate is proportional to the number of staff, which varies over time) as candidate predictor variables for Q-Lasso.

Motivated by operations management literature that identifies important drivers of wait time for an ED, the generalized fluid model estimators used in Q-Lasso account for other relevant “workload” in the hospital, in addition to the queue of patients waiting to start treatment, and for time variation in sequencing rules. For example, Saghafian et al. (2014) show how variation in the processing requirements for patients in the ED (notably by triage level) and sequencing rules impact wait times. Both Jaeker and Tucker (2012) and Shi et al. (2012) show that congestion in other parts of the hospital can “spill over” to increase overall waiting times in the ED, as some admitted patients are boarded in the ED due to physical or staffing constraints in the hospital.

Q-Lasso is also motivated by research that demonstrates the potential for statistical learning to improve health care systems, e.g., in the design of clinical trials (Bertsimas et al. 2013) and nurse staffing in an ED (Rudin and Vahn 2014). In particular, recent emergency medicine literature recommends predicting the ED wait time using quantile regression (Sun et al. 2012, Ding et al. 2010) and potentially providing a range (median to 90th percentile) for the wait time, because the median may be insufficiently accurate and patients dislike waiting longer than predicted. Those quantile regression methods neglect important predictor variables and structural relationships identified by the aforementioned operations management literature, though Sun et al. (2012) show that incorporating the queue length and recent processing rate as predictor variables improves accuracy.

A few other recent papers employ statistical learning in wait time prediction. For example, Balakrishna et al. (2014) and Simaiakis and Balakrishnan (2014) estimate aircraft taxi-out times (the time from gate departure to takeoff) using reinforcement learning and regression

trees, respectively. Senderovich et al. (2014) find that, in a call center, k-means clustering (predicting the wait time for a new arrival based on the average wait for customers that arrived while the queueing system was in a similar state, in the past) is less accurate than simply using the wait time of the last customer to enter service.

Providing wait time information can beneficially influence customer behavior, according to Armony and Maglaras (2004), Jouini et al. (2011), Allon et al. (2011) and literature surveyed in those papers. For example, in a call center model, providing an accurate wait time prediction and a “call back” option shapes arrivals so as to reduce customers’ mean and worst-case waits (Armony and Maglaras 2004). Empirical evidence from call centers suggests that providing the expected wait time for arriving customers increases the immediate abandonment (balking) rate when the system is congested (Mandelbaum and Zeltyn 2009, Yu et al. 2014), whereas providing simple information (that the wait time is low, medium or high) reduces abandonment rates for all levels of the message (Yu et al. 2014).

To beneficially influence patient behavior in the ED setting, Dong et al. (2015) call for *accurate* ED wait time prediction, in order to reduce overall waiting by balancing the load among nearby EDs, in a manner similar to the ambulance diversion studied by Deo and Gurvich (2011), Yu et al. (2014) and Xu and Chan (2014). By analyzing search engine data and the rolling average wait times on websites of 211 U.S. hospitals, Dong et al. conclude that patients increasingly are searching for ED wait times, and are using that information in choosing which local hospital’s ED to go to. Unfortunately, however, the rolling average wait times published by hospitals are not informative about the current state of the ED (Armony et al. 2012) so can increase overall waiting by causing patients to choose the wrong hospital (Dong et al. 2015). Another caveat is that a long predicted wait might cause people who should go to the ED to choose not to; thus, publishing the current expected wait time might increase a hospital’s profitability but reduce social welfare (Plambeck and Wang 2012).

Batt and Terwiesch (2015) conjecture that ED wait time prediction may reduce the rate at which patients leave without being seen. Empirically, in an ED that does not provide a wait time prediction, Batt and Terwiesch (2015) observe that to the extent that the ED is crowded, patients are more prone to leave without being seen. Patients might wrongly infer that the wait is long by focusing on the queue and not the processing rate, like the retail customers studied by Lu et al. (2013). Providing a wait time prediction may shorten the perceived wait and increase tolerance for waiting, according to literature surveyed by Jouini et al. (2011) and

Batt and Terwiesch (2015). Indeed, insights from the implementation of Q-Lasso at SMMC suggest that wait time prediction can improve patients’ experience of waiting and might deter some from leaving without being seen.

The remainder of the paper is structured as follows: §2 describes metrics for wait time, the four trial hospitals and the distinct data set used in wait time prediction for each hospital. §3 describes five candidate methods for ED wait time prediction: a best rolling average, two fluid model estimators, the method from Sun et al. (2012) and Q-Lasso. §4.1 evaluates the performance of these methods, and §4.2 identifies which candidate predictor variables are valuable in Q-Lasso, using historical data from the four trial hospitals. §5 provides insights from the implementation of Q-Lasso at SMMC: in particular, §5.1 reports post-implementation changes in patients’ experience of waiting and behavior; §5.2 reports the post-implementation accuracy of Q-Lasso; §5.3 and §5.4 describe challenges in implementation and practical ways to improve accuracy. Lastly, §6 highlights the main results and provides guidance for hospitals regarding ED wait time prediction.

2. Description of Hospitals, Data and Wait Time Metrics

This paper employs data sets from four hospitals. Three are private teaching hospitals located in New York City that did not provide permission to be named, and so are called Hospitals 1, 2 and 3. The fourth, SMMC, is a non-teaching, public hospital located in San Mateo County, California. As a public hospital, SMMC receives funding from the county and primarily serves its low-income residents. Only 8.5% of patients that visit the SMMC ED are admitted to the hospital versus an average of 18% for Hospitals 1, 2 and 3, possibly due to lower primary care utilization by patients of SMMC.

Figure 1 shows the basic process flow at all four hospitals’ EDs and the two metrics for wait time published at SMMC. The squares correspond to events recorded in the SMMC patient visit data. Most importantly for this paper, “start of treatment” is the time at which the patient first sees a *provider*, defined as a physician, physician assistant or nurse practitioner. The primary wait time metric considered in this paper is the *time-to-treatment*, defined as the time elapsed between registration and the start of treatment. The second metric published at SMMC is the remaining time-to-treatment for a patient at triage. (The start of treatment for a patient is recorded as the time that a provider electronically “signs up” to treat that patient, which might not reflect the patient’s perspective, as discussed in §5.3.)

Why focus on time-to-treatment? Although the U.S Centers for Medicare and Medicaid Services (CMS) requires hospitals to report historical averages for both the time-to-treatment and the time-to-discharge (wait time from registration to discharge), ACEP (2012) advocates focusing on time-to-treatment when predicting a wait time for patients. According to Boudreaux et al. (2000), patients are more concerned about time-to-treatment than time-to-discharge. Moreover, the authors observed that errors in predicting time-to-discharge would be unacceptably large (at least with the available data and methods considered in this paper) because a patient’s time-to-discharge is highly variable, depending on his idiosyncratic, sequential tests and procedures required for diagnosis and treatment, and the length of time that he must remain in the ED for observation after treatment.

In reporting the historical average time-to-treatment to CMS, SMMC and other hospitals exclude any waiting by a patient that leaves without being seen. Following that reporting standard and SMMC’s practices for internal performance evaluation, the authors exclude patients who left before the start of treatment in all calculations regarding time-to-treatment. Such patients comprise less than 4% of all visits to the four trial hospitals.

In the triage step shown in Figure 1, a nurse assigns a triage level depending on the patient’s acuity of condition and resource requirements. A triage level of 1 or 2 indicates that a provider should immediately evaluate the patient for a potentially life-threatening condition. The paper refers to patients of triage levels 1 and 2 as “high-acuity” and to patients of triage levels 3, 4 and 5 as “low-acuity.” Among low-acuity patients, a lower-numbered triage level indicates that the patient requires more resources from the ED and correspondingly imposes more work on the ED.

Regarding sequencing priority, a “Fast Track” is a small team of providers and nurses that focuses on low-acuity patients (primarily ones of triage levels 4 and 5) that can be treated quickly. SMMC operates a Fast Track on weekdays from 10am to 5pm, while Hospitals 1, 2 and 3 operate it on weekdays from 8am to 11pm. In deciding the order in which to treat patients, all other nurses and providers (and all staff, outside of Fast Track hours) prioritize high-acuity patients. In prioritizing among low-acuity patients, they consider the order in which the patients arrived, as well as patients’ conditions.

The SMMC data is more feature-rich than the data for the other three hospitals. As indicated by checkmarks in Table 1, the SMMC patient visit data contains the time of each patient’s registration, triage, chair or room assignment, start of treatment, disposition and

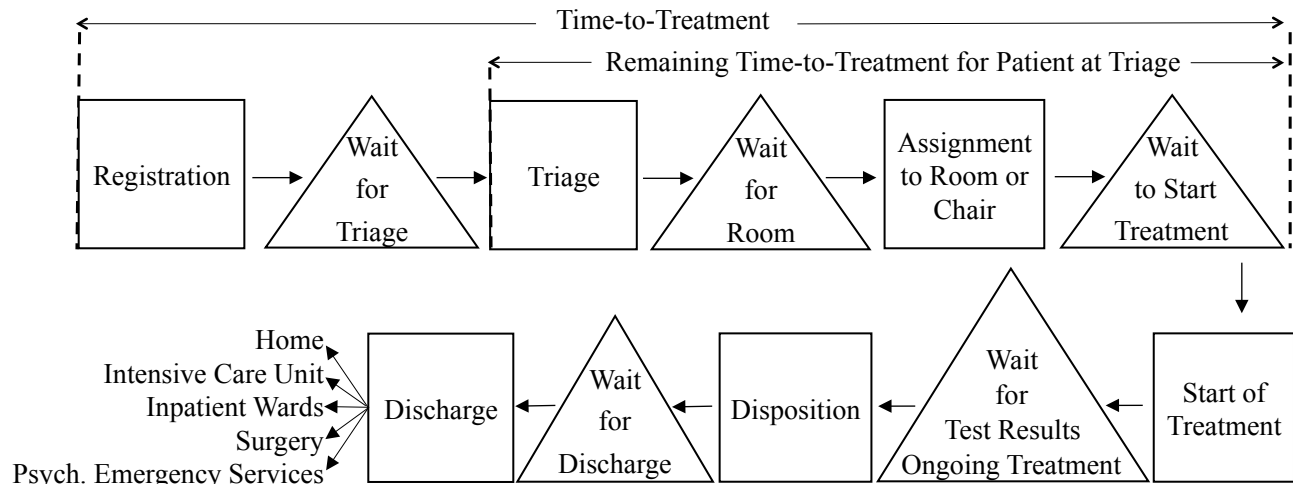


Figure 1 ED process flow diagram, showing the two metrics for wait time published at SMMC.

discharge. It also contains the patient’s mode of arrival, triage level, disposition (i.e., routing upon discharge), and the identification number of the provider that treated the patient. The SMMC patient visit data lacks the identification number for the nurse that served each patient, but SMMC provided a nurse staffing schedule.

Table 1 Each hospital’s data set has different features. “Required” features comprise the minimum data requirements for Q-Lasso and for a conventional rolling average method.

Features	SMMC	Hospital 1	Hospital 2	Hospital 3
Required				
Time of Arrival/Registration	✓	✓	✓	✓
Start Time of Treatment	✓	✓	✓	✓
Optional				
Time of Triage	✓			
Time of Rooming	✓			
Time of Disposition	✓			
Time of Discharge	✓	✓	✓	✓
Mode of Arrival	✓		✓	✓
Triage Level	✓	✓	✓	✓
Disposition Type	✓	✓	✓	✓
Provider Identifier	✓	✓	✓	✓
Nurse Identifier		✓	✓	✓
Nurse Staffing Schedule	✓			
Fast Track Schedule	✓	✓	✓	✓

Figure 2 shows the diurnal variation in patient arrival rates, numbers in system, wait times and staffing levels at SMMC. Compared to SMMC, Hospitals 1, 2 and 3 have qualitatively similar diurnal variation patterns, but the magnitude of diurnal variation in wait times is

smaller at Hospitals 2 and 3. Moreover, SMMC experiences the longest low-acuity wait times of the four, followed by Hospital 1; Hospital 2 and Hospital 3's wait times are much shorter. This may be explained in part by the fact that SMMC has the highest utilization, followed by Hospital 1: the ratio of mean hourly patient arrival rate to number of beds is 0.36, 0.21, 0.19 and 0.12 for SMMC and Hospitals 1, 2 and 3, respectively.

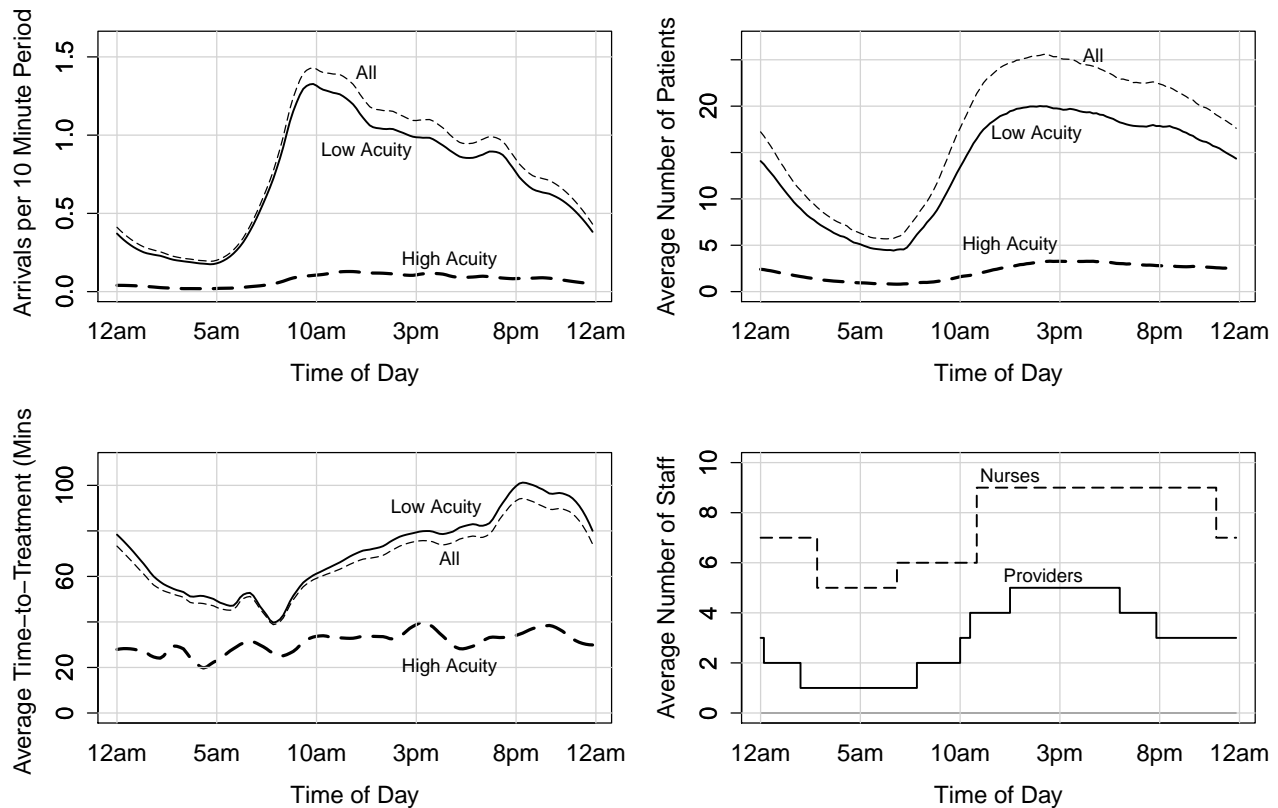


Figure 2 Top left, top right & bottom left panels: SMMC's average number of arriving patients, number of patients in the ED, and time-to-treatment, for all patients (dotted line), low-acuity patients (solid line) and high-acuity patients (dashed line), in each 10-minute period in the day. Bottom right panel: Number of nurses (dashed) and providers (solid) in the SMMC ED.

The data sets from the four hospitals vary in timespan and number of observations. Specifically, the data set from SMMC spans June 2012 to October 2013 and consists of 57,823 patient visits; the data set from Hospital 1 spans April 2011 to March 2012 and consists of 91,539 patient visits; and the data sets from Hospitals 2 and 3 span January to December 2011 and consist of 122,326 and 71,089 patient visits, respectively.

To implement and test methods for wait time prediction, the authors split each hospital's data set into two parts. The chronologically first 80% of patient visits comprise the *training*

set used to calculate the parameters required to implement each method. The remaining 20% of patient visits comprise the *test set* used to test the accuracy of each method.

3. Wait Time Prediction Methods

In contrast to the conventional practice of publishing a rolling average time-to-treatment for *all* patients, this paper focuses on predicting the time-to-treatment for a *low-acuity* patient. Hospitals should do so for two reasons. First, a wait time prediction is useful *only* for a low-acuity patient. Whereas a high-acuity patient must and will be treated quickly in any ED in the U.S., a low-acuity patient can beneficially use an accurate wait time prediction in choosing whether or not to go to an ED, when to go, or which ED to choose. Moreover, an accurate wait time prediction may decrease a low-acuity patient’s anxiety while waiting. Second, most of the patients that access and use a wait time prediction will likely be of low acuity: high-acuity patients are relatively rare (accounting for fewer than 10% of patient visits in each of the four trial hospitals) and many arrive at the ED on an ambulance, gurney/wheelchair or under custody of law-enforcement (47% for SMMC, 54% for Hospital 2 and 40% for Hospital 3; Hospital 1’s dataset does not include method-of-arrival). Third, the convention of publishing one estimate of the time-to-treatment for all patients leads to underestimation for low-acuity patients, the primary users of the estimate. Instead, a hospital should publish an estimate of the time-to-treatment for low-acuity patients, accompanied by a statement that patients with potentially life-threatening conditions will start treatment immediately. (Admittedly, some hospitals might prefer to underestimate the wait in order to attract patients, in conflict with this paper’s objective of improving prediction accuracy. Also, unlike the four trial hospitals, a Level 1 Trauma Center could have a majority of high-acuity patients. All the prediction methods in this section are applicable to a Level 1 Trauma Center, but their accuracy may be lower due to the uncertainty regarding how many high-acuity patients will arrive while a low-acuity patient waits to start treatment.)

A hospital will periodically “look up” new data from an EMR database and update the prediction. At SMMC, 10 minutes is the minimum update interval, because the “look up” requires 10 minutes. Using the historical data from all four trial hospitals, the authors confirmed that increasing the length of the update interval reduces the accuracy of Q-Lasso and rolling average methods. Therefore, the authors have implemented Q-Lasso at SMMC with a 10-minute update interval. For consistency, this paper also uses a 10-minute update interval in parameterizing all the methods proposed in this section, and in evaluating the accuracy of

those methods in §4 – that is, the error used to compute the MSE is the difference between a patient’s actual time-to-treatment and the predicted time-to-treatment during the 10-minute interval in which he or she registered at the ED.

3.1. Rolling Average

This paper identifies the “best” rolling average for each of the four trial hospitals. The first candidate form of rolling average is the average time-to-treatment among the last k low-acuity patients to start treatment. The second is the average time-to-treatment among low-acuity patients that started treatment within the last w hours. To identify which form and parameter to use for each hospital, the authors compute the MSE between a candidate rolling average, updated at the start of each 10-minute interval, and the time-to-treatment for the low-acuity patients that registered during each interval of the training data. The authors then vary the parameters k and w to identify the candidate rolling average and parameter level that achieves the minimum MSE. A fixed-time-window rolling average works best for all four trial hospitals; the best length for the time window w is 2 hours for SMMC, 4 hours for Hospital 1, and 6 hours for Hospitals 2 and 3. Apparently, the best length for the time window is inversely related to the degree of diurnal variation in wait times: to wit, wait times at Hospitals 2 and 3 vary less throughout the day than those at Hospital 1, which in turn experiences less variation than SMMC.

Commonly, hospitals publish a fixed-time-window rolling average without optimizing the length of the time window and without considering alternative prediction methods, though a few hospitals do publish a last- k -patients rolling average (Shen 2014, Dong et al. 2015). Hence the MSE for the best rolling average may be interpreted as a lower bound for the MSE in conventional practice.

3.2. Fluid Models

Following Ibrahim and Whitt (2009a), this paper considers the following two estimators based on simple fluid models. The first fluid model estimator for the time-to-treatment for a low acuity patient is

$$\frac{Q}{N\mu_{doc}} \quad (1)$$

where Q is the number of patients waiting in the ED to start treatment, N is the number of providers in the ED, and μ_{doc} is the rate at which a provider treats low-acuity patients. The second estimator is

$$\frac{Q}{\mu_{tot}} \quad (2)$$

where μ_{tot} is the total processing rate for low-acuity patients, which may vary periodically with the operating mode of the ED. Recall that high-acuity arrivals have preemptive priority to start treatment immediately, so Q will ordinarily be composed of low-acuity patients, and the processing rate parameters μ_{doc} and μ_{tot} will be small to the extent that the ED has a large arrival rate of high-acuity patients.

To calculate the fluid model estimators and evaluate their accuracy in §4.1, the number of providers in the ED, N , is calculated in the same manner as for Q-Lasso, described in the Q-Lasso subsection below. The processing rates μ_{doc} and μ_{tot} are calculated from the training set data for each hospital, using the naïve estimates

$$\hat{\mu}_{doc} = \frac{\sum_{t=1}^T M_t}{\tau \sum_{t=1}^T N_t}. \quad (3)$$

$$\hat{\mu}_{tot} = \frac{\sum_{t \in \mathbb{P}} M_t}{\tau \cdot |\mathbb{P}|}. \quad (4)$$

where $\tau = 10$ minutes, T is the number of 10-minute update intervals spanned by the training data, M_t is the number of low-acuity patients that started treatment in period t , N_t is the number of providers in the ED during interval t , and \mathbb{P} is the set of 10-minute intervals in the training data during which the ED operated in a particular mode (Fast Track on or off).

Whereas processing rate estimates (3) and (4) implicitly assume full utilization of providers or the ED, respectively, Simaiakis and Balakrishnan (2014, pp 14-15) estimate a processing rate by using a saturation plot to identify when full utilization occurs. Sun et al. (2012) estimate the ED processing rate by the number of patients that started treatment within the past hour, and use that as a predictor variable in their statistical learning method described in §3.3. §3.4 proposes an alternative, integrated approach to processing rate estimation and wait time prediction, Q-Lasso.

3.3. Quantile Regression Method from the Emergency Medicine Literature

In the emergency medicine literature, Sun et al. (2012) propose quantile regression to predict the ED wait time as a *linear* function of queue lengths and processing rates (in contrast to the *ratio* in the fluid model estimators (1) and (2)). They implement a system for providing wait time information to patients at triage. Therefore, they predict the wait from triage to start of treatment, contingent on a patient’s triage level. As predictor variables, they use the number of patients of each triage level that have been triaged and are waiting to start treatment, the number of patients of each triage level that started treatment within the past

hour (a notably different metric for processing rate than (3) or (4)), the time of day, and the day of the week.

Adapting the method of Sun et al. (2012) to predict the time-to-treatment for a low-acuity patient, this paper applies quantile regression (specifically choosing median-regression) in each hospital’s training data, using as predictor variables the number of patients in the ED waiting to start treatment, the number of low-acuity patients of each triage level that started treatment within the past hour, the time of day and the day of the week. Recall that the first of those predictor variables is the Q in fluid model estimators (1) and (2), and is ordinarily primarily composed of low-acuity patients.

As a second, alternative adaptation of Sun et al. (2012), the paper applies median-regression to predict the time-to-treatment for a low-acuity patient, using as predictor variables the number of patients in the ED that are waiting for triage, plus all the predictor variables used by Sun et al. (2012) (which are listed in the previous paragraph). The second adaptation is feasible only for SMMC, because the data sets for Hospitals 1, 2 and 3 lack triage times. Therefore, in attributing results to Sun et al. (2012) in §4, the paper uses the best-performing adaptation (with lowest test set MSE) for SMMC, and uses the first adaptation for Hospitals 1, 2 and 3.

3.4. Q-Lasso

Using the training set data, Q-Lasso constructs a linear “prediction function” to map information about the state of the ED into a prediction of the time-to-treatment for an arriving low-acuity patient. Recall that a hospital will update the state information and prediction periodically (for results in this paper, at intervals of 10 minutes). Therefore, the training set data must be formatted as $\{(w_1, \vec{P}_1), (w_2, \vec{P}_2) \dots (w_N, \vec{P}_N)\}$, where N is the number of low-acuity patient visits, w_i is the time-to-treatment for the i^{th} low-acuity patient, and $\vec{P}_i \in \mathbb{R}^M$ is the vector of values of candidate predictor variables at the start of the 10-minute interval in which patient i registered. In other words, \vec{P}_i represents the information about the state of the ED that could have been used to predict the time-to-treatment for patient i .

Lasso

The Lasso method selects a linear prediction function $f(P) = \vec{\beta}^\top \vec{P}$ to solve

$$\min_{\vec{\beta} \in \mathbb{R}^M} \frac{1}{N} \sum_{i=1}^N (w_i - \vec{\beta}^\top \vec{P}_i)^2 + \lambda \|\vec{\beta}\|_1. \quad (5)$$

The penalty term $\lambda \|\vec{\beta}\|_1$ prevents overfitting of the prediction function by forcing Lasso to use only a subset of the candidate predictor variables, i.e., to set the β coefficients of the others to zero. Indeed, that penalty term approach differentiates Lasso from classical variable selection methods in statistics, such as step-wise regression, which uses a discrete combinatorial approach to select subsets of candidate predictor variables. (The authors also experimented with stepwise linear regression, which generated less accurate predictions than Lasso despite being computationally very expensive.)

One may choose λ via cross validation. For all Lasso-based results reported in the paper, the authors split the training set into five equal subsets in chronological order, with no overlaps. Then, using the chronologically first 80% of patient visits in each subset, the authors chose $\vec{\beta}$ according to (5) for λ ranging from 0 to the point at which the MSE began to monotonically increase. For each value of λ , the authors computed the average MSE obtained over the latter 20% of visits in each subset. The optimal penalty λ^* was then the λ that obtained the lowest average MSE. For a more detailed overview of how to tune and implement Lasso, the reader should consult (Hastie et al. 2009).

Candidate Predictor Variables: Putting the “Q” in Q-Lasso

Q-Lasso incorporates the fluid model estimators (1) and (2) as candidate predictor variables. Q-Lasso also incorporates similar candidate predictor variables that – alone and in linear combination – generalize those fluid model estimators. Generalizing the fluid models described in §3.2, the time-to-treatment for an arriving low-acuity patient is:

$$\frac{\text{Workload}}{\text{Processing Rate}} \tag{6}$$

The workload is the work associated with all patients currently in the ED that must be completed before the ED begins to treat a newly arriving low-acuity patient. The workload depends on each patient’s processing requirements (captured imperfectly by triage level), the extent to which the ED has already processed the patient through the stages in Figure 1, and the sequencing priority for an arriving low-acuity patient. (Recall that during Fast Track hours, a subset of providers and nurses prioritizes low-acuity patients.) The processing rate in (6) depends on the numbers of providers and nurses in the ED and the sequencing priority rule. Therefore Q-Lasso constructs candidate predictor variables from the components represented in Table 2 by multiplying one entry from the first column with one entry from the second column, then dividing by one entry from the third column, and considering all

possible combinations. This yields $36 \times 4 \times 3 = 408$ candidate predictor variables for SMMC, $8 \times 4 \times 3 = 96$ for Hospital 1, and $10 \times 4 \times 3 = 120$ for Hospitals 2 and 3. Consistent with Table 1 in §2, SMMC has the most candidate components for the workload. In Table 2, the number of entries in each cell of the left column (different candidate workload components) is indicated in parentheses. Hospitals that have those candidate components for the workload are indicated in square brackets.

Table 2 Candidate predictor variables account for the workload, sequencing rule and processing capacity in the ED.

Workload	Sequencing	Processing Capacity
Number of patients in the ED in total (1) [SMMC, H1, H2, H3]	1 {Fast Track On}	Number of Providers
Number of patients registered but not yet triaged (1) [SMMC]		
Number of patients triaged but not yet assigned to a room/chair (1) [SMMC]		
Number of patients assigned to a room/chair that have not yet started treatment (1) [SMMC]	1 {Fast Track Off}	Number of Nurses
Number of patients in ED that have started treatment, in total or of triage level i , for $i \in \{1,2,3,4,5\}$ (6) [SMMC, H1, H2, H3]		
Number of patients that arrived by amb./fire dept./police/gurney but have not yet started treatment (1) [SMMC, H2, H3]		
Number of patients that arrived by amb./fire dept./police/gurney and have started treatment (1) [SMMC, H2, H3]		
Number of patients that have registered but not yet started treatment (1) [SMMC, H1, H2, H3]		
Number of patients that have been triaged but not yet started treatment, in total or of triage level i , for $i \in \{3,4,5\}$ (4) [SMMC]		1
Number of patients that have started treatment but not yet received disposition, in total or of triage level i , for $i \in \{1,2,3,4,5\}$ (6) [SMMC]		
Number of patients that have received disposition but not yet discharged in total or by disposition i , for $i \in \{\text{Intensive Care, Inpatient, Psych Emergency, Surgery}\}$ (5) [SMMC]		
Number of patients that have been triaged but not yet discharged in total or by triage level i , for $i \in \{1,2,3,4,5\}$ (6) [SMMC]	1	

Q-Lasso also includes the number of providers and number of nurses in the ED as simple candidate predictor variables. A hospital could calculate those numbers from a work schedule. However, the schedule may change, and the actual number of staff in the ED often deviates from the schedule due to flexible breaks, absenteeism, staff called in due to absenteeism or a surge in patient arrivals, etc. Therefore, this paper proposes an alternative approach. The number of providers in the ED is calculated as the maximum of two terms. The first term is the number of unique provider identifiers associated with patients that started treatment within the past 60 minutes, i.e., the number of unique providers that signed up to treat a patient within the past 60 minutes. The second term is the minimum number of providers

that must be in the ED according to standard operating procedure. (The choice of 60 minutes was motivated by the authors’ observation of the SMMC ED, wherein only rarely does a provider in the ED go for 60 minutes without signing up for a patient. However, such delays do occur, albeit infrequently, due to a lack of patient arrivals – typically late at night while the ED is operating with the minimum number of providers.) For the results reported in §4, the authors apply this dynamic approach to calculate the number of providers in the ED for all four trial hospitals, and to calculate the number of nurses in the ED for Hospitals 1, 2 and 3. Since the patient visit data from SMMC lacks a nurse identifier, the authors instead use the work schedule to impute the number of nurses in the SMMC ED.

Another candidate predictor variable is the best rolling average prediction, calculated as specified in §3.1 at the start of the 10-minute period in which the patient registers. A hospital that has an alternative prediction method (e.g., a favored rolling average method, or a prediction based on a sophisticated queuing analysis or simulation model) could incorporate that prediction as a candidate predictor variable.

In addition, one may construct three sets of deterministic candidate predictor variables belonging to \vec{P}_i using only a patient’s arrival time. The first set uses 143 binary variables to capture the 10-minute period of the day in which the patient registers. (In general, a hospital intending to update wait time predictions every m minutes would analogously construct $[(24 \times 60)/m] - 1$ binary period-of-day variables, with all variables taking the value of 0 except for the one corresponding to a patient’s registration period.) Second and similarly, one can represent the day of the week on which each patient arrives using 6 binary variables. Third, a binary variable can represent whether or not Fast Track was in operation when the patient registered. (Fast Track information is not captured by the period-of-day variables for the four hospitals in this paper, because these hospitals do not operate Fast Track on weekends.)

Optionally, one may construct candidate predictor variables using external, publicly available data sets. For example, SMMC staff suggested that weather impacts the arrival process, especially because many prospective patients must walk or wait outside to use public transport. Therefore, for the results reported in §4, the authors included the local average temperature in the previous 2 hours and a binary variable indicating whether or not rain or fog occurred in the last 2 hours as candidate predictor variables for each hospital. The authors also included the number of flu cases reported during the past week in each hospital’s metropolitan area, extracted from Google’s Flu Trend database.

The Online Supplement lists the “chosen” candidate predictor variables – i.e., those with non-zero coefficients in the prediction function – which are remarkably different for the four trial hospitals.

Q-Lasso is easy to implement with the free, open-source software R, using the **glmnet** package for Lasso and the example code in the Online Supplement to transform patient visit data from a conventional EMR system into the candidate predictor variables in Table 2.

Applying Q-Lasso to Alternative Wait Time Metrics or Groups of Patients

To apply Q-Lasso to predict the remaining time-to-treatment at triage for a low-acuity patient, one takes w_i in (5) to be the wait time from triage to the start of treatment for the i^{th} low-acuity patient in the training set, and $\vec{P}_i \in \mathbb{R}^M$ the vector of values of candidate predictor variables at the start of the 10-minute interval in which patient i was triaged. To apply Q-Lasso to predict the remaining time-to-treatment at triage for a patient of triage level t , one restricts the set $\{(w_1, \vec{P}_1), (w_2, \vec{P}_2) \dots (w_N, \vec{P}_N)\}$ to patients of triage level t , so that N is the number of visits by patients of triage level t in the training set. In the same manner, one may apply Q-Lasso to predict alternative metrics for wait time for alternative groups of patients.

4. Results in Test Data for the Four Hospitals

4.1. Performance of Various Methods in Predicting Time-to-Treatment

For all four hospitals, Q-Lasso is more accurate than all the other methods described in §3; see Table 3. Though all methods – including Q-Lasso – make some large errors, Q-Lasso achieves the greater accuracy largely by mitigating errors of underestimation, i.e., protecting patients from waiting much longer than predicted; see Table 4.

Table 3 Test set mean squared error (and standard error) of various methods for predicting time-to-treatment of low-acuity patients for each hospital. Errors are in minutes.

Prediction Method	Mean Squared Error			
	SMMC	Hospital 1	Hospital 2	Hospital 3
Best Rolling Average	2517.2 (73.7)	2725.6 (33.0)	970.1 (22.0)	551.1 (18.2)
Fluid (1)	2658.9 (79.6)	2779.8 (39.7)	1004.9 (22.2)	602.1 (19.5)
Fluid (2)	2491.1 (79.1)	2428.2 (33.7)	961.3 (21.6)	567.5 (19.2)
Sun et al. (2012)	1854.7 (66.3)	2456.6 (35.5)	954.5 (24.1)	525.6 (19.5)
Q-Lasso	1693.4 (56.2)	2056.2 (26.4)	864.8 (20.8)	480.4 (17.1)

Most importantly (because hospitals commonly publish a rolling average), Q-Lasso is *much* more accurate than even the best rolling average method. To visualize how Q-Lasso outperforms a rolling average, consider Figure 3, which shows mean wait times and predictions made by the best rolling average and Q-Lasso by time of day (at SMMC and Hospital 1). Due to the diurnal fluctuation in wait times and the lag in the rolling average, the rolling average tends to overestimate when the wait time is short, and underestimate when the wait time is long. In particular, a rolling average tends to underestimate the wait during the period of the day in which most low-acuity patients arrive (10am to midnight), often substantially. This fact is corroborated by the prediction error quantiles in Table 4. At SMMC, for example, for 5% of low-acuity patients, the time-to-treatment is more than 1.5 hours longer than the best rolling average prediction. A conventional rolling average would be even less accurate and more prone to underestimation, because hospitals typically include high acuity patients' near-zero waits and do not parametrically optimize the rolling average, as documented in §3. In contrast, Figure 3 shows that Q-Lasso traces the diurnal fluctuation in the time-to-treatment much better than a rolling average. Table 4 shows that Q-Lasso mitigates the problem of underestimation, in that the magnitude of the error with Q-Lasso tends to be smaller than that with a rolling average among patients with underestimated wait times. Although Table 4 focuses on SMMC, the error quantiles for Hospitals 1, 2 and 3 are qualitatively similar and lead to the same insight, though Hospitals 2 and 3 have weaker diurnal variation in wait times and correspondingly smaller errors with a rolling average.

Relative to the best rolling average method, Q-Lasso reduces the MSE by 33.3%, 25.1%, 12.0% and 13.2% for SMMC and Hospitals 1, 2 and 3, respectively. The relatively greater percentage improvement in accuracy over a rolling average at SMMC stems from three factors. The first factor – high utilization and associated high variability in wait times – helps to explain the high percentage improvement for SMMC and Hospital 1. Wait times exhibit higher variance at SMMC and Hospital 1 (3996 and 2150, respectively) than at Hospitals 2 and 3 (1095 and 694, respectively). The greater the variation in wait times, the greater the opportunity for Q-Lasso to explain the variation through candidate predictor variables that “capture” events and conditions in the ED. For example, consider a major car pile-up that causes multiple high-acuity patients to enter the ED, increasing the time-to-treatment for low-acuity patients (especially at SMMC, operating at extremely high utilization). Q-Lasso can account for such shocks within 10 minutes through the workload predictor variables.

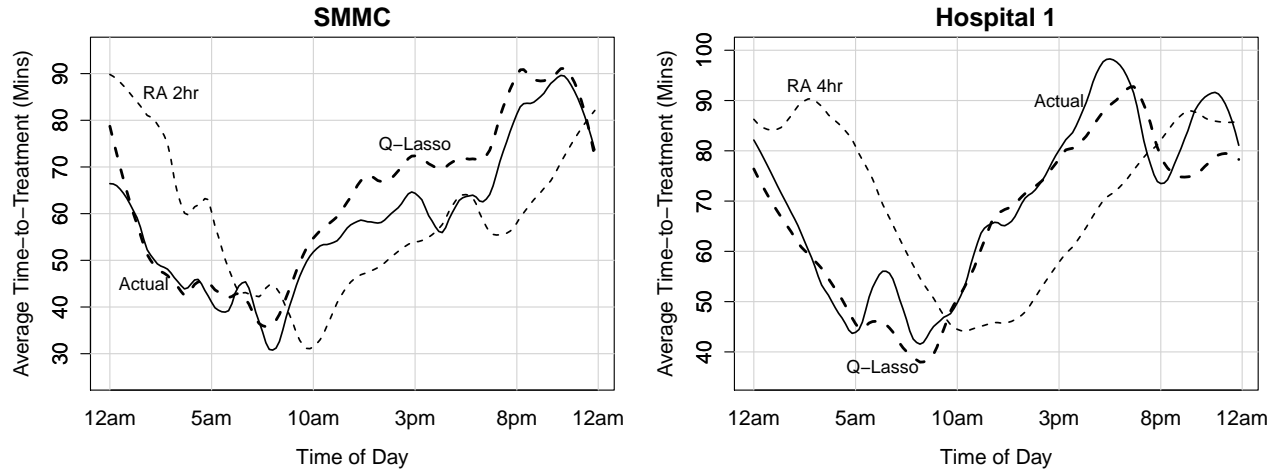


Figure 3 Average, for low-acuity patients arriving within each 10-minute interval of the day, of the time-to-treatment (solid line), predicted time-to-treatment with Q-Lasso (thick dashed line) and the best rolling average (thin dashed line) for each hospital.

Furthermore, reconsidering the diurnal variation in wait times shown in Figure 3, recall that a rolling average is inaccurate to the extent that the diurnal variation in wait times is large. SMMC and Hospital 1 experience greater diurnal variation in patient wait times than Hospitals 2 and 3, so Q-Lasso has relatively more “room” for improvement over a rolling average at SMMC and Hospital 1.

Table 4 Error (predicted minus actual time-to-treatment) quantiles of SMMC test set. Errors are in minutes.

Prediction Method	Error Quantiles (mins)				
	5%	25%	50%	75%	95%
Best Rolling Average	-91.7	-28.6	-2.4	17.5	63.6
Fluid (1)	-108.2	-41.4	-15.0	1.5	33.8
Fluid (2)	-104.4	-41.2	-15.9	-0.5	26.3
Sun et al. (2012)	-73.8	-18.5	4.6	18.8	51.2
Q-Lasso	-61.5	-10.1	9.4	27.2	61.8

A second reason that SMMC achieves the greatest improvement with Q-Lasso over a rolling average method is that the SMMC data is the most feature-rich, providing Q-Lasso with more candidate predictor variables and hence the ability to construct a more accurate prediction function. A third reason is that Hospitals 1, 2 and 3 made changes in ED operating procedures during the data collection period and, for the results reported in this section, the Q-Lasso prediction function is not updated to account for such changes. §5.3 proposes a better training procedure for Q-Lasso, which enables Q-Lasso to adapt to operational changes in the ED.

Fluid Model (2) (sequencing-rule-dependent service rate) is more accurate than Fluid Model (1) (staff-dependent service rate) for all four trial hospitals, as is evident from Table 3. This echoes the observation, explained in §4.2, that the various fluid model estimators that account for Fast Track (constructed from the first two columns of Table 2) are valuable candidate predictor variables for Q-Lasso, whereas those based on the number of providers or number of nurses (constructed using the third column of Table 1) are not.

Q-Lasso outperforms the superior Fluid Model (2) by 32.5%, 15.9%, 11.2% and 15.6% for SMMC and Hospitals 1, 2 and 3 respectively. Q-Lasso does so by simultaneously optimizing the parameters and assigning weights to a wide variety of candidate fluid model estimators, notably ones that account for detailed information about the workload in the ED, whereas Fluid Models (1) and (2) account only for the queue of patients waiting to start treatment, and have naïve processing rate estimates. When that queue is empty, the estimated time-to-treatment with Fluid Models (1) and (2) is 0, whereas in reality a patient must be triaged and roomed and, in some cases, wait for an x-ray or other test before seeing a provider. Q-Lasso accounts for that minimal delay by inclusion of an intercept. That helps to mitigate the problem, evident in Table 4, that Fluid Models (1) and (2) tend to underestimate the time-to-treatment.

In comparison with Sun et al. (2012), Q-Lasso reduces the MSE by 8.7%, 12.7%, 9.4% and 7.3% respectively for SMMC and Hospitals 1, 2 and 3. Importantly for patients, Q-Lasso mitigates the extremes of underestimation by Sun et al. (2012) evident in Table 4. Any median-based regression is prone to underestimation because the time-to-treatment distribution is right-skewed (the median is lower than the mean), so predictions that minimize deviations from the median are lower than predictions that minimize deviations from the mean.

4.2. Which Candidate Predictor Variables are Valuable in Q-Lasso?

A hospital might achieve comparable prediction accuracy with greater ease of implementation by considering only a few of the proposed candidate predictor variables. Therefore, the authors evaluate the MSE for Q-Lasso with only the easiest-to-implement candidate predictor variables, and then, sequentially, with incorporation of increasingly-difficult-to-implement sets of candidate predictor variables.

Determining whether or not a set of candidate predictor variables is “valuable” (yields a statistically significant improvement in the Q-Lasso MSE at the 5% level) is technically

challenging. The Q-Lasso prediction with a restricted set of candidate predictor variables may be highly correlated with the Q-Lasso prediction with incorporation of additional candidate predictor variables. Hence a small improvement in the MSE may be statistically significant, even if the two approaches' MSEs exhibit large or seemingly overlapping confidence intervals. The authors establish statistical significance using a modified pairwise t-test that accounts for such correlation. The method is described in detail in the Online Supplement.

Table 5 reports the Q-Lasso MSE with the easiest-to-implement set of candidate predictor variables in the top row and, in subsequent rows, reports the MSE with the incorporation of each increasingly-difficult-to-implement set of candidate predictor variables. Each column, corresponding to one of the four trial hospitals, must be read from top to bottom. Within each column, the first “*” indicates that the MSE is statistically significantly smaller than the MSE in the top row; each subsequent “*” indicates that the MSE is statistically significantly smaller than the MSE in the previous row with a “*”. In short, a set of candidate predictor variables is valuable if a “*” appears in the row in which it is introduced.

The sets of candidate predictor variables are defined, and denoted in Table 5, as follows: the set of *deterministic* period-of-day and day-of-week variables (**D**); one variable, the number of low-acuity patients in *queue* to start treatment (**Q**); the set of *workload* variables listed in the first column of Table 2 (**W**); the set of *sequencing* priority variables (**S**), which includes the indicator for Fast Track and includes all the candidate predictor variables from Table 2 that employ an indicator for Fast Track On or Off from the second column, and a “1” from the third column; one variable, the best *rolling* average (**R**), optimized for each hospital as described in §3.1; the set of all variables involving the *number* of providers or *number* of nurses (**N**), i.e., the *staff-count* variables estimated as described in §3.4, as well as all the *staff-based* candidate predictor variables in Table 2 (those that do not have a “1” in the third column); and “**All**” is all of the candidate predictor variables proposed in §3.4, including ones based on data from external sources: flu trends, temperature and weather information.

Table 5 shows that although the deterministic day-of-week and period-of-day variables (**D**) alone perform remarkably well, incorporating each of queue length (**Q**), workload (**W**), sequencing priority variables (**S**), rolling average (**R**) and staff-based variables (**N**) statistically significantly reduces the MSE at one or more hospitals. On the other hand, incorporating the flu, weather and temperature variables (**All**) does *not* significantly reduce the MSE.

Table 5 Test set mean squared error (and standard error) for Q-Lasso with increasingly large subsets of the candidate predictor variables. Errors are in minutes.

Candidate Predictor Variables	Mean Squared Error			
	SMMC	Hospital 1	Hospital 2	Hospital 3
D	2714.9 (61.2)	2698.4 (36.3)	942.2 (20.7)	530.3 (16.8)
D, Q	2402.3 (61.1)*	2119.7 (28.7)*	865.6 (20.5)*	492.5 (17.2)*
D, Q, W	1739.3 (56.3)*	2064.3 (26.9)*	860.2 (20.4)*	486.7 (17.0)*
D, Q, W, S	1742.7 (56.4)	2058.5 (26.7)*	859.5 (20.6)	483.6 (17.1)*
D, Q, W, S, R	1725.3 (56.9)	2037.2 (26.1)*	862.5 (20.7)	481.6 (17.2)*
D, Q, W, S, R, N	1696.1 (55.9)*	2042.4 (26.2)	862.7 (20.8)	480.2 (17.1)
All	1693.4 (56.2)	2056.2 (26.4)	864.8 (20.8)	480.4 (17.1)

The deterministic Q-Lasso prediction (Q-Lasso using only period-of-day and day-of-week variables) is more accurate than the best rolling average method at Hospitals 1, 2 and 3. (Compare the MSEs in row (**D**) of Table 5 with those for the best rolling average in Table 3 above.) This insight is important because, as described in detail in §5.3, the greatest difficulty in implementing a wait time prediction system is to set up the “live data” interface, by which the wait time prediction is updated dynamically based on the state of the ED. A hospital can avoid this difficulty and still potentially reap a benefit by implementing the deterministic Q-Lasso prediction.

A second important insight from Table 5 is that the number of patients waiting to start treatment (**Q**) is valuable but inadequate: at all four trial hospitals, further incorporating the workload associated with patients that have already started treatment (**W**) significantly improves prediction accuracy. Moreover, at two of the trial hospitals, accounting for sequencing priority also improves prediction accuracy. Incorporating sequencing priority variables (**S**) in addition to the workload variables (**W**) enables Q-Lasso to assign different coefficients in the prediction function to elements of the workload, depending on whether or not the ED is operating in Fast Track. Intuitively, this enables Q-Lasso to account for the dependency of the workload and processing rate on the sequencing priority rule.

In contrast, the number of providers and number of nurses are only valuable as predictor variables at one hospital (SMMC), either alone or combined with workload and sequencing rule predictor variables according to Table 2. This can be seen from the limited improvement in MSE across Hospitals 1, 2 and 3 from adding the staff-based predictor variables (**N**). Moreover, so long as the time-of-day and day-of-week variables are incorporated ahead of the

staff-count variables, this result is robust to the order in which variables are incorporated to the model. This suggests that the period-of-day and day-of-week variables (**D**) capture the diurnal variation in staff counts at Hospitals 1, 2 and 3, but not at SMMC, perhaps because the proposed method for calculating staff counts from staff identifiers in the patient visit data is imperfect. Alternatively, the functional forms adopted in Fluid Model (2) and in Table 2 may simply be inappropriate for Hospitals 1, 2 and 3’s EDs, due to the complex interaction of patients, providers and nurses in the ED. For example, if staff speed up when the workload per staff member is large, time-to-treatment would be concave rather than linear in the number of staff (Shunko et al. 2014).

Finally, whereas the above analysis assumes that the deterministic time-of-day and day-of-week variables (**D**) are easiest to implement, a hospital that has a rolling average wait time prediction system before implementing Q-Lasso might instead begin the sequential variable incorporation process with that rolling average. Therefore the authors pursued an analogous analysis, starting with (**R**), then incorporating the other candidate predictor variables in the same order as in Table 5. Incorporating only the deterministic variables (**D**) significantly lowered the MSE for all four hospitals. Additionally including the simple workload candidate predictor variables (**Q** and **W**) also significantly reduced all four MSEs, consistent with the previous analysis. In contrast, inclusion of sequencing priority variables (**S**) significantly reduced the MSE only for Hospital 3 – at odds with the previous analysis, where it also reduced the MSE at Hospital 1. As in the previous analysis, however, inclusion of staff-based variables (**N**) significantly reduced the MSE only for SMMC. No additional candidate predictor variables yielded a significant reduction in MSE.

5. Implementation of the Wait Time Prediction System at SMMC

Since February 14, 2015, SMMC has been publishing the predicted time-to-treatment for low-acuity patients on an external website and publishing the predicted *remaining* time-to-treatment (wait from triage to start of treatment) on a screen in the triage room. Each is calculated using Q-Lasso. Readers may access the external website from SMMC’s homepage, <http://www.sanmateomedicalcenter.org/>, and the live feed to the triage room screen at http://204.114.51.208/bytrriage_all.php.

5.1. Patients' Behavior and Experience of Waiting

During the first three months of implementation, few patients (only 3 to 6 unique users per day) checked the ED wait time on the SMMC external website. SMMC plans to advertise the wait time website to all its existing patients to increase awareness and usage.

In contrast, nearly 10,000 low-acuity patients could see the predicted wait time on the screen in the triage room, during just the first three months of implementation. Hence the triage room setup is useful for studying how wait time information impacts patients' experience of waiting and their behavior.

SMMC staff report that providing the wait time prediction has improved patients' waiting experience. The charge nurse stated, "Patients became happier after we started showing the predicted wait times to them because now they have a better sense of how long they have to wait." She and other staff also report that patients have become less anxious while waiting to start treatment. According to the SMMC chief medical information officer, also a practicing physician in the ED, "setting an expectation about the wait time empowers patients." The caveat is that, in cases where the actual wait exceeds the prediction, patients complain or ask the staff why they must continue waiting. Therefore, although the staff are pleased to have the wait time screen in triage, some have asked the authors to "fix the problem" that the system underestimates the wait for some patients.

This qualitative improvement in patients' waiting experience might yield measurable changes in patient behavior. In particular, one of SMMC's most important performance metrics – the percentage of patient that leave without being seen (LWBS) – significantly decreased from 3.6% in the month preceding implementation (January 12 to February 13, 2015) to 2.8% in the first month after implementation (February 14 to March 15, 2015). Though LWBS rates have subsequently risen back above 3%, they remain lower than recent pre-implementation rates, suggesting that the provision of wait time information may have some mitigating effect. However, though preliminary analysis tentatively supports this association (see the Online Supplement for details), rigorously determining whether or not such an effect exists will require a controlled, randomized study – an intriguing direction for future research.

5.2. Post-Implementation Accuracy

Results from the first three months of implementation show that Q-Lasso is much more accurate than a rolling average method. For low-acuity patients who visited the SMMC ED between February 14 and May 15, 2015, Q-Lasso achieved a 33% lower MSE in predicting

the time-to-treatment on the external website than would have occurred with the best (2-hour-window) rolling average method. The Q-Lasso MSE was 1120.7 with a standard error of 25.0, whereas the 2-hour-window rolling average MSE was 1663.7 with a standard error of 39.1. In the triage room, Q-Lasso achieved a 30% lower MSE in predicting the residual time from triage to treatment than would have occurred with the 2-hour-window rolling average. The Q-Lasso MSE was 998.6 with a standard error of 24.0, whereas the rolling average MSE was 1429.4 with a standard error of 35.9.

The Q-Lasso prediction of the remaining time-to-treatment at triage was 10.9% more accurate than the Q-Lasso prediction of time-to-treatment on the external website. This improvement reflects that for a low-acuity patient, the wait between registration and triage is substantial and variable (having a mean and standard deviation of 14 and 10 minutes, respectively). The state of the ED can change substantially while a patient waits for triage, so updating the state information substantially improves the prediction of the remaining time-to-treatment. As the external website is updated to reflect state information, one might predict the remaining time-to-treatment for a low-acuity patient at triage simply by subtracting a constant (mean wait from registration to triage) from the Q-Lasso prediction of time-to-treatment on the external website; for SMMC, the MSE from doing so was only 4% larger than the Q-Lasso prediction of remaining time-to-treatment.

5.3. Challenges of Implementation

As operating practices and conditions change over time, a hospital may improve prediction accuracy by “retraining” – updating Q-Lasso’s coefficients using recent training data. The authors used the training data from 2012 and 2013 for the first month of the Q-Lasso implementation because more recent data was unavailable. In that first month, from February 14 to March 15, 2015, Q-Lasso nevertheless outperformed the best rolling average by 32% for the time-to-treatment on the external website, and 30% for the remaining time-to-treatment at triage. Thereafter, retraining using newly available recent data – specifically, data spanning December 21, 2014 to February 13, 2015 – yielded the 33% and 30% improvement in performance reported in §5.2. The retraining improved accuracy: between March 16 and May 15, the “old” version of Q-Lasso based on the training data from 2012 and 2013 would have outperformed the best rolling average by just 22% and 24% for the website and triage room screen, respectively. The Online Supplement shows that frequent retraining improves

accuracy, in SMMC’s historical data, and explains how a hospital can easily automate the retraining process.

The most challenging step in the implementation of Q-Lasso is to create the “live data” interface between the EMR system and the server that performs the wait time predictions using the Q-Lasso prediction functions. Creating this interface was particularly difficult at SMMC because a third-party vendor provides the software for the EMR system. The vendor was unwilling to share their proprietary software, and the particular form of application programming interface required to support the wait time prediction system did not exist. Therefore the authors created and installed an entirely new interface program, written in C#, which required guidance and coordination from both the vendor and SMMC staff.

Another challenge is that the EMR data might not represent a patient’s actual experience. For example, by monitoring the implementation at SMMC, the authors observed a systematic bias in the EMR data that tends to lead to underestimation, from a patient’s perspective, of the time-to-treatment. Specifically, after “signing up” to treat a patient, a provider typically collects charts, reviews test results, takes notes, walks to the bedside, etc. before starting to treat that patient. Anecdotally, especially at times when the ED is extremely busy, a provider may “sign up” simultaneously for multiple patients, then proceed to see the patients sequentially. The authors observed and manually recorded the delay from “sign up” until a low-acuity patient sees the provider, by conducting a preliminary motion study at SMMC for 5 hours on March 10 and 12. The delay ranged from less than 1 minute to 18 minutes, with a mean of 4 minutes, for the 13 low-acuity patients that started treatment during the observation period. From the patient’s perspective, the accuracy of Q-Lasso could potentially be improved by increasing the predicted time-to-treatment to account for the delay from “sign up” until a patient actually sees the provider. That delay could be larger than observed in the initial motion study at SMMC, especially in other hospitals that employ time-to-treatment-based performance metrics that incentivize providers to sign up for patients prematurely (Shen 2014).

Furthermore, the EMR data might not represent a patient’s actual condition. For example, staff may initially assign a patient a low-acuity triage level and later upgrade the patient to a high-acuity triage level or vice versa. The EMR record for each patient contains only the most recent triage level which, in historical data, corresponds to the patient’s final triage level. Such reassignment occurs in only 1% of patient visits to the SMMC ED, so is unlikely to

substantially impact the accuracy of Q-Lasso, but a hospital with more frequent triage-level reassignment should be aware of this potential inconsistency.

5.4. Individualizing the Prediction at Triage

The authors analyzed and rejected the option to provide a separate prediction – for patients of each triage level 3, 4 and 5 – of the remaining time from triage to start of treatment. §3.4 describes how to do so with Q-Lasso. Doing so would fail to significantly improve prediction accuracy. Based on the post-implementation data from SMMC, separate prediction would increase the MSE for patients of triage levels 4 and 5 by 4.8% and 1.3%, respectively, reduce the MSE for patients of triage level 3 by 5%, and reduce the MSE for all low-acuity patients by only 1%. One rationale for the poor performance of this approach is that partitioning the patient visit data by triage level results in having too little data to estimate the prediction function effectively for each triage level, especially for patients of triage level 5 which account for only 6% of all low-acuity patients.

The practical reason that SMMC shows only one prediction on the screen in triage (for low-acuity patients) is that showing three different wait times (for patients of each triage level 3, 4 and 5) would require the triage nurse to explain the assigned triage level to each low-acuity patient and would likely frustrate some patients. In particular, at times that SMMC operates Fast Track, the predicted wait for level 5 patients is often much lower than that for patients of triage levels 3 or 4, which would likely motivate the latter to perceive their wait times as unfairly long or demand an explanation from the triage nurse.

Conceivably in the future, as a substitute for the screen showing one wait time prediction, the triage nurse could give an individualized wait time prediction to each patient verbally. Thus, Q-Lasso prediction accuracy could be further improved by incorporating the patient's chief complaint, tests ordered and other information gathered at triage as candidate predictor variables, in addition to triage level. The triage nurse could input this information into his or her computer, which would return the individualized predicted wait time for the patient via a real-time interface with the server performing the wait time predictions.

Q-Lasso could be applied to predict a patient's time from triage to discharge, using the aforementioned additional candidate predictor variables. However, insofar as accuracy in predicting time-to-discharge remains low, a hospital that wishes to give patients information about the time-to-discharge should use an alternative method, such as quantile regression, to generate a range instead of a single-number prediction.

6. Concluding Remarks

This paper introduces the Q-Lasso method for wait time prediction in a complex, data-rich service system. Q-Lasso inputs many candidate predictor variables – including ones inspired by candidate fluid models for the system – to the Lasso method of statistical learning. Lasso chooses which of the candidate predictor variables to use, and how to predict the wait time as a linear function of those chosen predictor variables, with the objective of minimizing the mean squared error in prediction. Intuitively, in assigning coefficients to the various fluid-model-inspired candidate predictor variables, Q-Lasso automatically decides how best to estimate the current workload and processing rate in the system. Moreover, in assigning coefficients to all the candidate predictor variables, Q-Lasso automatically decides how best to predict the wait time as a weighted combination of fluid model estimators, a rolling average estimator, deterministic estimators reflecting the diurnal variation in wait time, and other potentially relevant information.

In predicting the emergency department wait time-to-treatment for four different hospitals, Q-Lasso is more accurate than rolling average methods (currently used by hospitals), stand-alone fluid model estimators (from the operations management literature), and quantile regression methods (from the emergency medicine literature). Furthermore, Q-Lasso improves accuracy largely by reducing errors of underestimation, in which a patient waits for longer than predicted. The improvement relative to a rolling average is greatest for hospitals with the highest utilization, with the greatest random and diurnal variability in wait times, and with the most feature-rich data. Specifically, having data on the numbers of patients of various triage levels at various stages of processing in the ED, and accounting for the time-varying sequencing priority rule significantly improve the accuracy of Q-Lasso.

At SMMC, a Q-Lasso prediction system provides live wait time information to prospective patients via an external website, and to current patients via a screen in the triage room. Insight gained through the SMMC implementation of Q-Lasso provides guidance for other hospitals. First, hospitals should provide a wait time prediction to patients at triage – in addition to or instead of doing so on an external website. Wait time prediction at triage at SMMC improves patients' experience of waiting and might reduce the rate at which they leave without being seen. Whereas relatively few patients may access the external website, all low-acuity patients benefit from prediction at triage. Prediction accuracy is better at triage, even without basing the prediction on the information about the patient's condition gathered

at triage. Providing a wait time prediction at triage instead of on an external website would also mitigate an important ethical and liability concern. As with some instances of heart attack or stroke, a patient might have misleadingly mild symptoms and not realize that the condition is potentially life-threatening. Seeing a long wait time on an external website might dissuade such a patient from seeking immediate treatment. At triage, a triage nurse directs such patients into immediate treatment, so providing wait time information for low-acuity patients at triage poses little or no risk for such high-acuity patients.

The second piece of guidance pertains to updating the wait time prediction. The main difficulty in implementing a conventional rolling average or Q-Lasso is to establish the “live data” interface to pull information about the state of the ED from the EMR system. If a hospital is unable or unwilling to set up the “live data” interface, the hospital can nevertheless give patients a deterministic time-varying wait time prediction that can be significantly more accurate than a conventional rolling average. A hospital that already publishes a rolling average can easily implement Q-Lasso with free, open-source software and the example code in the Online Supplement. For greatest accuracy, the wait time prediction should be updated, via the “live data” interface, as frequently as possible (at 10-minute intervals at SMMC). Over months, as operating practices and conditions change substantially, a hospital should update Q-Lasso using recent training data.

Third, hospitals should evaluate the consistency of data recorded in the EMR system with patients’ experiences. In particular, the start of treatment recorded in the EMR system may be earlier than the time that a patient first sees a provider. Uncorrected, that biases a conventional rolling average and Q-Lasso toward underestimating a patient’s time-to-treatment.

Fourth, accuracy is not necessarily the right objective. According to the feedback from SMMC nurses and patients, the objective should reflect that patients become frustrated when they wait for longer than predicted, and are happy when the wait is shorter than predicted. The objective should also reflect how wait time prediction potentially shapes demand (patient arrivals and abandonment behavior) and thus impacts patients’ wait times and welfare, and hospitals’ revenues. A hospital could implement Q-Lasso with an alternative objective, by substituting that objective for the mean squared error in (5). However, that would require creation of specialized code, because the open source software for Lasso assumes the MSE objective.

Acknowledgements

The authors are extremely grateful to Melissa Valentine and Stefanos Zenios for encouragement and assistance in obtaining data; to Mor Armony, Bob Batt, Jérémie Gallien, Diwakar Gupta, Alan Scheller-Wolf, Pengyi Shi, Qiong Wang, Galit Yom-Tov, Jim Dai and our *great* anonymous review team for guidance and insights; and to Lisa Behraves, Julie Griffiths, Kim Huynh, Syed Khan and the rest of the staff at SMMC ED for their patience, practicality and support.

In addition, M. Bayati acknowledges the support of NSF awards CMMI:1451037 and CCF:1216011.

References

- ACEP. 2012. Publishing wait times for emergency department care: An information paper. *American College of Emergency Physicians* .
- Allon, G., A. Bassamboo, I. Gurvich. 2011. We will be right with you: Managing customer expectations with vague promises and cheap talk. *Operations Research* **59(6)** 1382–1394.
- Armony, M., S. Israelit, A. Mandelbaum, Y.N. Marmor, Y. Tseytlin, G.B. Yom-Tov. 2012. Patient flow in hospitals: A data-based queueing-science perspective. *Working Paper* .
- Armony, M., C. Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Operations Research* **52(4)** 527–545.
- Armony, M., N. Shimkin, W. Whitt. 2009. Impact of delay announcements in many-server queues with abandonment. *Operations Research* 66–81.
- Balakrishna, P., R. Ganesan, L. Sherry, B.S. Levy. 2014. Estimating taxi-out times with a reinforcement learning algorithm. *IEEE/AIAA 27th 3.D.3-1 – 3.D.3-12*.
- Batt, R.J., C. Terwiesch. 2015. Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science* **61 (1)** 39–59.
- Belloni, A., V. Chernozhukov, Ch. Hansen. 2014. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81** 608–650.
- Bertsimas, D., A. O’Hair, S. Relyea, J. Silberholz. 2013. An analytics approach to designing clinical trials for cancer. *Working Paper* .
- Boudreaux, E.D., R.D. Ary, C.V. Mandry, B. McCabe. 2000. Determinants of patient satisfaction in a large, municipal ed: The role of demographic variables, visit characteristics, and patient perceptions. *The American Journal of Emergency Medicine* **18(4)** 394–400.
- Deo, S., I. Gurvich. 2011. Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science* **57(7)** 1300–1319.

- Ding, R., M.L. McCarthy, J.S. Desmond, J.S. Lee, D. Aronsky, S.L. Zeger. 2010. Characterizing waiting room time, treatment time, and boarding time in the emergency department using quantile regression. *Academic Emergency Medicine* **17** 813–823.
- Domingos, P. 2012. A few useful things to know about machine learning. *Communications of the ACM* **55** 78–87.
- Dong, J., E. Yom-Tov, G.B. Yom-Tov. 2015. The impact of delay announcements on hospital network coordination and waiting times. *Working Paper* .
- Hastie, T., R. Tibshirani, J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.
- Ibrahim, R., M. Armony, A. Bassamboo. 2014. Does the past predict the future? the case of delay announcements in service systems. *Working Paper* .
- Ibrahim, R., W. Whitt. 2009a. Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management* **11** 397–415.
- Ibrahim, R., W. Whitt. 2009b. Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Science* **55** 1729–1742.
- Ibrahim, R., W. Whitt. 2011a. Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations Research* **59(5)** 1106–1118.
- Ibrahim, R., W. Whitt. 2011b. Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals. *Production and Operations Management* **20(5)** 654–667.
- Jaeker, J.A., A.L. Tucker. 2012. Hurry up and wait: Differential impacts of congestion, bottleneck pressure, and predictability on patient length of stay. *Harvard Business School Working Paper* .
- Javanmard, A., A. Montanari. 2014. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning* Forthcoming.
- Jouini, O., Z. Aksin, Y. Dallery. 2011. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management* **13** 534–548.
- Lu, Y., A. Musalem, M. Olivares, A. Schilkret. 2013. Measuring the effect of queues on customer purchases. *Management Science* **59(8)**.
- Mandelbaum, A., S. Zeltyn. 2009. Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research* **57(5)** 1189–1205.
- Plambeck, E., Q. Wang. 2012. Implications of hyperbolic discounting for optimal pricing and scheduling of unpleasant services that generate future benefits. *Stanford Working Paper* .
- Rudin, C., G. Vahn. 2014. The big data news vendor: Practical insights from machine learning analysis. *Working Paper* .

- 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.
- Senderovich, A., M. Weidlich, A. Gal, A. Mandelbaum. 2014. Queue mining-predicting delays in service processes. *Advanced Information Systems Engineering*. **Springer** 42–57.
- Shen, S. 2014. Personal conversation between the authors and sam shen, medical director (stanford school of medicine, division of emergency medicine). June 9, 2014.
- Shi, P., M.C. Chou, J.G. Dai, D. Ding, J. Sim. 2012. Hospital inpatient operations: Mathematical models and managerial insights. *Working Paper* .
- Shunko, M., J. Niederhoff, Y. Rosokha. 2014. Humans are not machines: Impact of queueing design on service time. *Working Paper* .
- Simaiakis, I., H. Balakrishnan. 2014. A queueing model of the airport departure process. *Transportation Science* Forthcoming.
- Sun, Y., K.L. Teow, C.K. Ooi B.H. Heng, S.Y. Tay. 2012. Real-time prediction of waiting time in the emergency department, using quantile regression. *Annals of Emergency Medicine* **60(3)** 299–308.
- Xu, K., C.W. Chan. 2014. Using future information to reduce waiting times in the emergency department. *Working Paper* .
- Yu, Q., G. Allon, A. Bassamboo. 2014. How do delay announcements shape customer behavior. *Management Science* Forthcoming.