# The Impact of Delay Announcements on Hospital Network Coordination and Waiting Times

Jing Dong, Galit B. Yom-Tov, Elad Yom-Tov

# The Impact of Delay Announcements on Hospital Network Coordination and Waiting Times

**(Authors' names blinded for peer review)**

We investigate the impact of delay announcements on the coordination within hospital networks using a combination of empirical observations and numerical experiments. We show that patients take delay information into account when choosing emergency service providers and that such information can help increase coordination in the network, leading to improvements in performance of the network, as measured by Emergency Department wait times. Our numerical results indicate that the level of coordination that can be achieved is limited by the patients' sensitivity to waiting, the load of the system, the heterogeneity among hospitals, and, importantly, the method hospital use to estimate delays. We show that delay estimators that are based on historical average may cause oscillation in the system and lead to higher average waiting times when patients are sensitive to delay. We provide empirical evidence which suggests that such oscillations occurs in hospital networks in the US.

*Key words*: Delay Announcements, Emergency Department, Queueing Network Coordination, Join the Shortest Queue, Pooling, Cost of Waiting

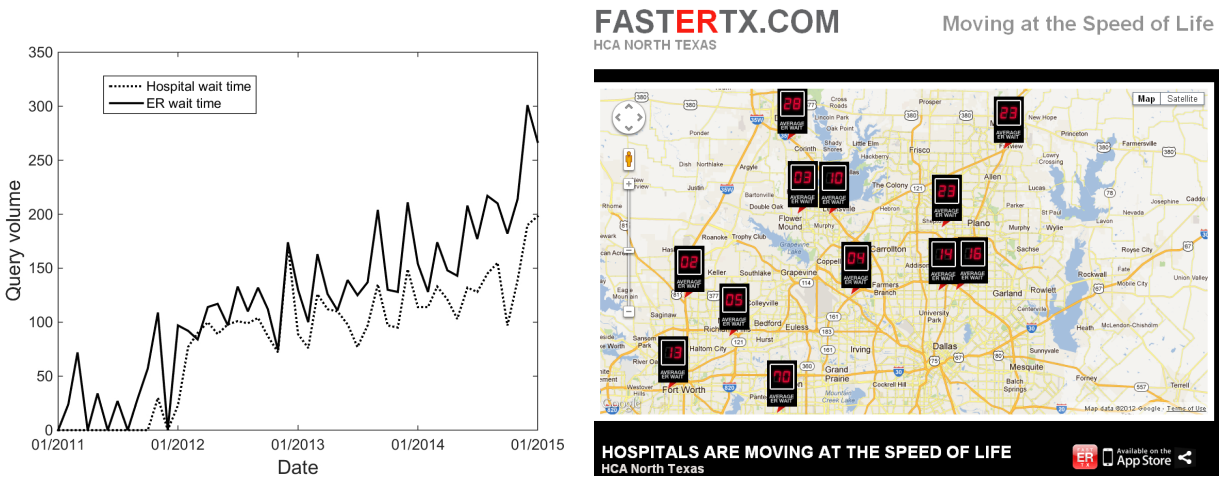
## 1. Introduction

Delay announcements have become commonplace in service systems. They serve as a means to influence quality perception, system operation, and customer sentiment towards the service provider. As a consequence, understanding the impact of delay announcements on customer choices and system operations, as well as the development of methods to support such announcements, has attracted the attention of the Operations Research and Management communities in the past few years. Thus far, most of the research in this area concentrated on call center announcements, where it was shown that such information influences, for example, customer abandonment (Mandelbaum and Zeltyn 2013, Yu et al. 2014).

In recent years, a growing number of hospitals have begun posting their Emergency Department (ED) waiting times on websites, billboards and smartphone apps (see for example Figure 1(b)). Here we examine the effect of this information (also referred to as delay announcements or waiting time announcements) on the choices made by patients. Although the primary consideration of patients in selecting an ED is its timely provision of treatment, it is clearly not the only consideration. The reputation of the hospital, its expertise, and the patients' medical insurance plan, as well as recommendation by the primary physician all influence such choice (Marco et al. 2012). Given the effort required to provide waiting time information for hospital ED services, it is important to ask, do people actually seek this information when choosing a hospital? Is the proportion of people who seek such information large enough to have an operational impact on the healthcare system, and hospital networks in particular? Do hospitals provide the right information to help achieve coordination (resource pooling) in the network?

Here we attempt to answer questions regarding the publication of wait times and their impact on the customer choice and the system performance using a combination of empirical analysis and numerical experiments. The empirical data provides objective evidence for the use of delay information. We base our analysis on two primary sources of data: first, we collected information of real ED delay announcements from more than 200 hospitals in the US over a period of 3 months; second, we observe anonymized queries made to the Bing search engine by people seeking ED delay information during that period. The latter information is used to estimate public interests in such delay announcements and their influence on the delays themselves. We notice that a growing number of people search for ED wait times: Figure 1(a) shows the query volume to the Google search engine for "hospital wait time" and "ER wait time" (trends.google.com). As the figure demonstrates, this volume has been steadily rising in the past 4 years.

Using insights from queueing theory, we conduct empirical analysis of the hospital networks, and study the operational influence of delay announcements on future delays and on correlation of delays among hospitals. Throughout the paper we refer to two hospitals with correlated waiting




(a) Google Trends query volume regarding ED wait times

(b) Real time delays announcement web-page of emergency departments in north Texas, USA.

**Figure 1** **Internet publication and query trends of ED wait times.**

times as synchronized EDs and measure that level of *synchronization*. We then use a stylized simulation model to investigate how system characteristics such as patients' sensitivity to waiting, load and different delay estimators influence the phenomena we observed in the data.

### 1.1. Scientific Background

Delay announcements have a measurable influence on customer satisfaction (Carmon and Kahneman 1996, Larson 1987). Such announcements can be given in a variety of ways, ranging from vague information on the current load to specific information on the customer's location in the queue or their expected waiting times. The effect of these messages differ. Munichor and Rafaeli (2007) showed that, in a call center environment, providing information about the location in the queue will result in lower abandonment rates and higher customer satisfaction, as compared to other waiting time fillers such as music or apologies. Allon et al. (2011) developed a game theoretic model of a strategic service provider and a strategic customer, which provides a theoretical basis for determining how vague delay announcements should be.

One of the challenges in incorporating detailed delay announcements is producing a *credible* estimation of the delay. In a series of papers, Ibrahim and Whitt proposed several delay estimators, based on queueing theory, for customers joining a multi-server service system. They considered queuing systems with a time-homogeneous arrival process and staffing level, as well as a time-varying one (Ibrahim and Whitt 2009, 2011). Their proposed estimators are based on a real-time history of the queue, e.g., the delay of the last customer who entered the agent's service (LES) and the total current delay of the customer at the head of the line (HOL). These estimators perform well in reality as was shown in Senderovich et al. (2014). Senderovich et al. (2014) used queue mining

techniques to solve the on-line delay prediction problem, validating the queueing theory-based predictors with real data.

Estimating delays in EDs is much more difficult than in call centers, because of the inherent complexity and transient nature of these systems. ED patients do not wait in a single queue but instead have a process they must go through. This process involves multiple resources (physicians, nurses, labs, etc.), and takes between 3 to 6 hours to complete. The apparent complexity is even greater when one considers the fact that the arrival rate is time-varying, and that the route of each patient is unknown ahead of time. Plambeck et al. (2014) developed a forecasting method of ED delay estimator that are based on a combination of queueing and machine learning methods. In our data, none of the hospitals uses such sophisticated models. Instead, hospitals publish historic average waiting times using a 4-hour moving average, a measure which has become the convention in US hospitals.

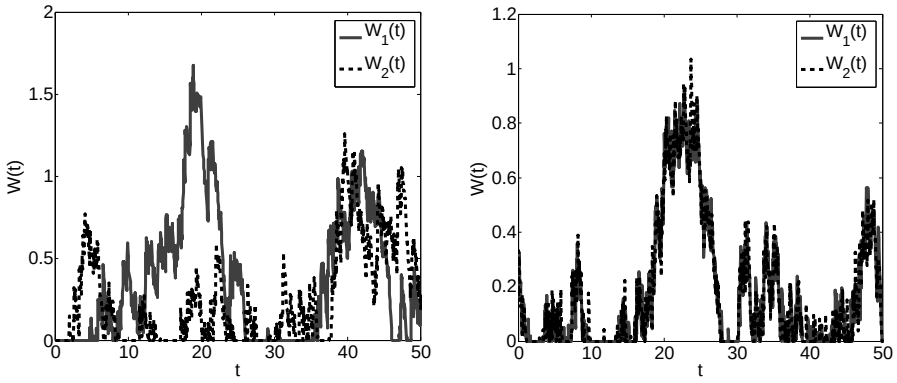
Delay announcements influence not only customer satisfaction but also customer actions. Announcing the expected delay as customers enter the system, especially in heavily loaded periods of time, may cause customers to balk (leave the system upon arrival) or abandon after a short time (Mandelbaum and Zeltyn 2013, Yu et al. 2014). Delay announcements may serve as a means to reduce the offered load on a service system. If announcing a specific (long) wait causes some customers to abandon, the waiting time of the remaining customers will be shortened. This feedback may cause difficulties in analyzing the steady-state performance of the system. Armony et al. (2009) explored the effect of delay announcement in an $M/GI/s+GI$ queue, proposing two approximations for the steady state performance of such systems. Ibrahim et al. (2013) developed delay estimators that took these feedback effects into account.

Since delay announcements influence customer actions and customer waiting costs (Yu et al. 2014), they can be used as an operational tool. For example, delay announcements are used in call centers to help customers choose their time of service via a call-back option (Armony and Maglaras 2004). In theme parks, delay announcements assist customers in choosing preferred queues, hence raising questions on their impact on resource allocation (Kostami and Ward 2009).

The above-mentioned research investigated the impact of delay announcements on the company which provides the information, not from the network perspective thereof. When multiple companies are present (for example, when several EDs are located in the same area), it is important to investigate the impact of such announcements on social welfare in a network setting, where the announcements of one service provider may impact the demand for services at other providers. Moreover, in the setting of an ED, service providers are not only competing with each other but also have incentives to cooperate. On one hand, hospitals want patients to come to their ED, as the EDs are considered as the 'gateway' to a range of services offered by a hospital. On the other

hand, the resources (capacity) of a hospital are very limited, due to the expensive nature of such services, which in turn causes occasional high congestion and long waiting times.

When an ED is overcrowded, delays may cause quality of care to deteriorate (Chalfin et al. 2007). At such times the hospital has an incentive to ease some of the load by reducing the arrival rate. Some hospitals do this through ambulance diversions. However, ambulance diversion may turn away patients who are most at need of medical care, whereas hospitals may intend to reduce the arrival rate of the least severe patients. Hence, hospital have an incentive to announce delay information of the ED as a means to influence the behavior of the least severe patients, especially when the ED is crowded. These patients can then chose a different ED or delay their visit. In general, non-acute patient population can account for up to 90% of ED visits (Plambeck et al. 2014). Assuming patients take such information into account, announcements are expected to smooth the demand for hospital services throughout the day (and week) and balance patient load between nearby hospitals. Figure 2 compares the simulated sample path for two uncorrelated systems, where patients randomly choose which hospital to attend, and two correlated (fully synchronized) systems, where patients always choose the hospital with the shortest waiting time. We base our analysis on the observation that in the latter, the waiting times of the two hospitals are synchronized. Therefore, in this paper, we identify connections between delay announcements and correlation among workloads at geographically proximate EDs.




(a) Two independent (uncorrelated) queues

(b) Two fully synchronized (correlated) queues

**Figure 2 Unsynchronized v.s. synchronized systems.**

From a theoretical point of view, hospitals are akin to independent agents where there is no social planner that balances the load between them. It is known that if some coordination exists, for example, by having customers always Joining the Shortest Queue (JSQ), the efficiency of the network could be improved to be almost like that of a fully pooled system (Foley and McDonald

2001). Even when the fraction of customers choosing a server by the JSQ policy is small, this policy is still advantageous (Turner 2000). Hence, one of the operational rationales for providing delay announcements is that they might improve the efficiency of the hospital network, even if not all patients are strategic or sensitive to waiting.

### 1.2. Goals and main contributions

We make the following main contributions:

- We show that patients explore delay information provided by hospitals, and do so at a growing rate.
- Customers take delay information into account when deciding on ED providers and there are enough of them to have a pooling effect on the hospital network. This is evident as the number of delay reporting hospitals per unit area and the number of queries about them are significant indicators in explaining synchronization levels in hospital networks. The distance between the hospitals is another factor that influence synchronization levels. The larger the distance between two hospitals, the lower the effect they have on one another.
- Our numerical results show that the synchronization level between two hospitals are influenced by how sensitive customers are to delay, the load of the system, and the heterogeneity between hospitals in terms of customer preferences and size.
- If the appropriate method is used for delay estimation (and loads are fairly balanced), the social welfare of patients in this network increases, i.e., delay announcements will decrease the average waiting times of *all* hospitals in the network.
- The accuracy of the waiting time estimator as well as the delay of the estimator in reflecting the true waiting time have a profound impact on the effectiveness of delay announcements. We find that the commonly used method of a 4-hour moving average could be problematic. This method can cause the load to oscillate between hospitals when customers are very sensitive to delay. Our empirical analysis supports this finding and suggests that small differences in wait times between geographically proximate hospitals translate, on average, to large future differences and vice versa.

## 2. Empirical study of announcement impact

In this section, we use the data described above to find objective indicators that patients indeed use delay information in choosing an emergency service provider, and this has a profound influence on the network coordination. We also explore how sensitive patients are to differences between waiting times of different hospitals.

### 2.1. Data sources used in this research

First, we identified 211 US hospitals which published their waiting times using RSS feeds as of March 2013. We collected these waiting times every 5 minutes by polling their RSS feeds between March and June 2013 (inclusive). The time provided is from entering the ED to when one is expected to see the first medical provider. No real time information on length-of-stay, or classification according to severity is provided. All these hospitals use the same delay estimators, namely, a moving average over a 4-hour time window. According to the explanation in those website, the wait time provided does not apply to urgent patients, as these patients are prioritized according to their medical conditions. Hence, hospitals urge patient to ignore this information if they are in an immediate threat to their lives. From our data it appears that this information is updated every 15 minutes.

To estimate the total number of hospitals in each area we obtained a list of 36,438 hospitals identified in the Bing local search application.

Finally, we extracted all queries made using the Bing search engine between March and June 2013 (inclusive) which resulted in visits to the web pages of one of the hospitals in our list. Each query contained an anonymized user identifier, time stamp, user location (GPS information for mobile users and zip-code information for other users), the query text, and the pages which were clicked as a result of this query.

### 2.2. Results

As a preliminary step, we clustered the 211 hospitals according to their wait times into three groups using the K-means algorithm. The resulting clusters partition hospitals into three categories: low wait, medium wait and high wait times. Figure 3 shows the average waiting time of the three groups over a 3-week period, in 5-minute resolution, which demonstrate both the daily pattern and the variation in waiting time among the three groups.

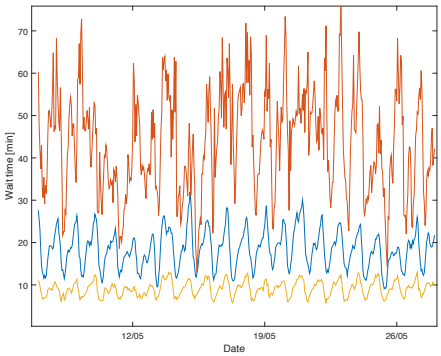


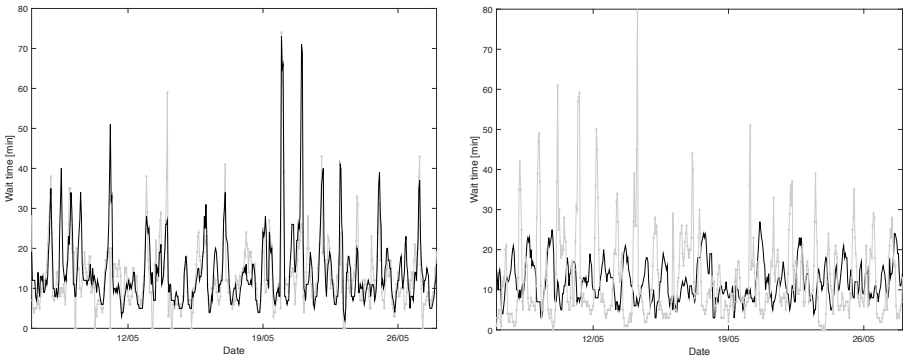

**Figure 3** **Average waiting time of the three groups of hospitals—low wait, medium wait and high wait times.**

In most of the empirical analysis, we want to ignore synchronization that is due to diurnal patterns. Hence, for each hospital, we removed the hourly and daily waiting time trends from the observed waiting times in the following manner. First, the average waiting times for each hospital at each hour of the day and each day of the week was computed. Then, for each hospital, a linear predictor was trained to predict the waiting times using the appropriate hourly and daily waiting times, and the predicted waiting times were removed from the observed waiting times. We refer to the resulting detrended waiting times as the Residual Waiting Times (RWT).[1]

Figure 4 plots the RWT of two hospital pairs, suggesting that some ED pairs are more synchronized than others. The first factor to influence synchronization is distance—we expect geographically close hospitals to be more synchronized. Indeed, there is a negative correlation between the distance of hospitals and the level of synchronization of these hospitals. Specifically, the correlation between the distance of hospitals and the correlation of their RWT is $-0.138$ (p-value $< 10^{-5}$).

Observation 1. Distance influences synchronization levels: synchronization is stronger for closer hospitals. In other words, the larger the distance between two hospitals, the lower the effect they have on one another.

This geographical synchronization could be attributed to several factors, e.g., information provided to ambulance services. What we seek to understand is whether the announcement of anticipated delays for the general population contributes to this synchronization, and to what extent.




(a) Highly synchronized EDs (32 km apart) (b) Poorly synchronized EDs (1257 km apart)

**Figure 4 Waiting time announcements of different EDs.**

We note in passing that wealth does not seem to be associated with more hospitals publishing their wait times: To each of the 36,438 hospitals identified in the Bing local search application we

[1] The average ratio between the weight of the hourly trend and the weight of the daily trend was 0.999, indicating that both terms had an almost identical effect. Together, these two variables explained 85.0% of the variance of the original waiting times.

found the 2012 adjusted gross income data of the closest zip code from IRS data. Using these data we found that wait times are published in locations where the median income is the same to within 0.5% ($P = 0.017$, ranksum test).

**2.2.1. Do patients use delay information when deciding on ED services?** As noted above, people query for delay information at a growing rate in the past several years. To investigate whether patients use delay information and to understand its influence on synchronization, we model the level of synchronization between clusters of geographically close hospitals (less than 20 km apart) using five variables.

To define 'geographically close hospitals', we clustered the hospitals which published wait times according to their geographic location by performing Agglomerative Hierarchical Clustering (Duda et al. 2001) until no hospitals were found within 20km. This resulted in 46 clusters.

Then, we trained a regression tree (see Figure 5) to predict the average correlation of RWT within a cluster and modeled each cluster using the following variables:

1. Number of EDs reporting wait time per square km within a cluster.
2. Number of hospitals per square km (either reporting wait time or not) within a cluster.
3. Number of queries made (to the Bing search engine) about the reporting hospitals within the cluster.
4. Average wait time of the hospitals within the cluster.
5. Number of pediatric EDs within the cluster (some EDs do not, or only during part of the day, cater to children).

Applying the leave-one-out cross-validation method, we estimated the performance of the model and found that the Spearman correlation between predicted and actual average RWT correlation was $\rho = 0.397$ ($P = 0.006$), indicating that the variables provide a good explanatory power for the dependent variable, e.g., the synchronization level as measured by the correlation of RWT.

Observation 2. Customers seem to take delay information into account in their decision making process: As evident from Figure 5, the number of delay reporting hospitals per unit area and the number of queries are significant indicators to explain the hospital network synchronization. The higher the number of reporting hospitals per unit area is, the higher the correlations are observed. More queries about the waiting time are associated with higher synchronization levels, as is the existence of a pediatric ward.

**2.2.2. How sensitive patient are to waiting time differences?** In this section, we investigate the cost-of-waiting. The cost-of-waiting influences the sensitivity of patients to waiting times. If this cost is high, even small gaps of waiting times between geographically close hospitals will

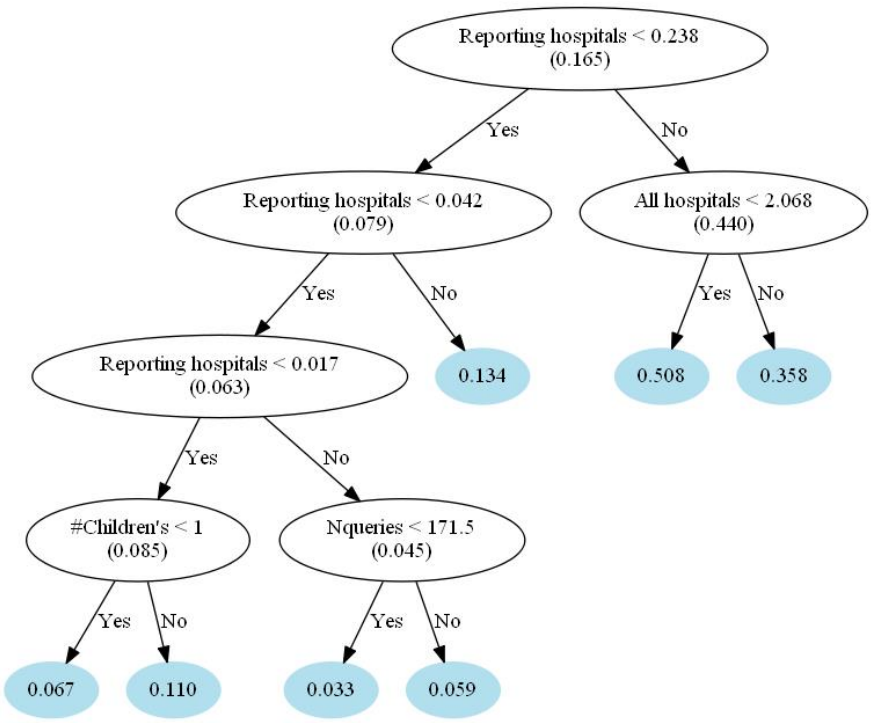



**Figure 5** **Regression tree classifier for predicting the in-cluster RWT correlation. Decision nodes show the splitting variable and the average correlation at the node (in parenthesis). Leaf nodes show the average correlation at the node.**

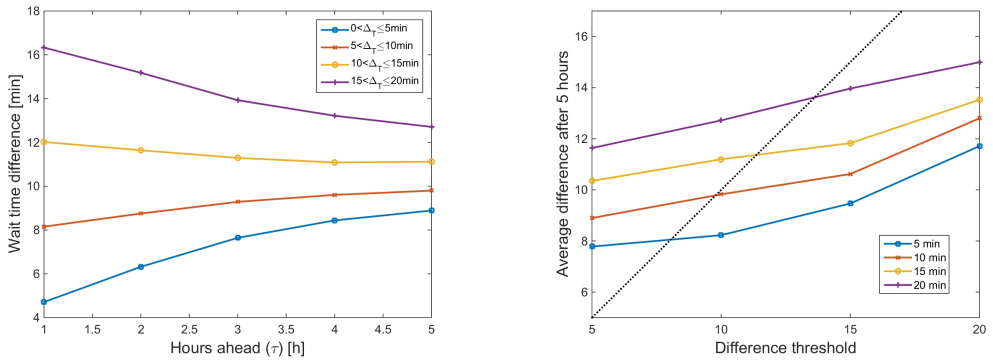
influence the patient's choice. Naturally, a patient with higher severity has a higher cost of waiting. Note that the most urgent patients are usually transported by ambulance to the nearest ED, and are prioritized in the queue, hence, the reported delays in the website do not apply to them. Therefore, we examine the impact of delay announcements on the non-urgent population, that may check the delays before choosing an ED.

We computed the average future differences in waiting times between adjacent hospitals as a function of the current differences. Let $W_i(T)$ denote the waiting time of hospital $i$ at time $T$. First, we calculated the difference in waiting times $\Delta_T = W_i(T) - W_j(T)$ at each time instance $T$ between pairs of hospitals that are separated by 20km or less. We further calculate $\Omega_T = min(W_i(T), W_j(T))$, which is the lower of the waiting times among the two hospitals. Then, we computed the average difference of waiting times between each pair at time $T + \tau$, where $\tau = [1, 2, 3, 4, 5]$, and stratified by $\Delta_T$ at 5-minute increments, that is, $0 < \Delta_T \leq 5$ up to $15 < \Delta_T \leq 20$ minutes.

The results of this calculation are shown in Figure 6. Figure 6(a) demonstrates that smaller initial differences in waiting times are associated with higher future differences, and vice versa. We hypothesize that this counter-intuitive result is due to the cost of waiting, coupled with the method by which waiting times are given, i.e., as an average of the past 4 hours of wait times. A

linear cost of waiting (such as the one we use in Section 3) implies that if the difference in wait times is small, patients choose hospitals randomly, and consequently differences tend to grow. If differences are large, patients will tend to choose the less loaded one and the difference will become smaller. As we show in Section 3.3, as the cost of waiting increases, such differences are less prone to occur. In addition, we show in Section 3.2 that the use of the 4-hour estimators can also cause such phenomena, even when the cost of waiting is high.

Figure 6(b) shows that future time differences are also dependent on the actual waiting times at time $T$, not only the differences: The figure shows $\Delta_{T+\tau}$ for $\tau = 5$ as a function of two parameters, namely, $\Delta_T$ and $\Omega_T$. As the figure demonstrates, when the initial difference is small ($< 10$ min) the difference in wait times after 5 hours increases, whereas when it is high ($> 10$ min) it decreases. However, the rate of change is strongly dependent on the shorter of the wait times: When this time is long (i.e., 20 min), future differences in wait times change linearly. In contrast, if the minimal wait time is low (i.e., 5 min), the change is much greater when initial differences are large compared to when they are small.




(a) Average future wait times as a function of the difference between the wait times of pairs of hospitals.

(b) Average $\Delta_{T+\tau}$ for $\tau = 5$ as a function of the difference between the wait times at $T$, for four levels of $\omega_T$. Dotted diagonal line indicates $\Delta_T = \Delta_{T+\tau}$.

**Figure 6 Average future wait times.**

We attempted to model the future difference between wait times of adjacent hospital pairs using the current difference and the number of queries made about both hospitals in the current time period. That is, given the difference in wait times $\Delta_T$ between two hospitals at time $T$ and the number of queries made about these hospitals, $Nq$, between $T$ and $T+1$ hours, we predict $\Delta_{T+\tau}/\Delta_T$ where $\tau = [1, 2, 3, 4]$ using a linear regression model:

$$\Delta_{T+\tau}/\Delta_T = w_1 \cdot \Delta_T + w_2 \cdot Nq_T + w_3$$

We then analyzed the correlation between the parameters of the models and the model coefficients, using $\Delta_T$ values between 0 and 20 minutes, at 2 minute intervals, that is, $0 \le \Delta_T < 2$, $2 \le \Delta_T < 4$, (time units in minutes), etc.. Specifically, the $R^2$ between the middle of each $\Delta_T$ range (e.g., 1min for the 0–2min range) and the weight in the regression model of the actual $\Delta_T$ ($w_1$) is 0.38, while the $R^2$ between the minimal $\Delta_T$ and the weight of the queries ($w_2$) is 0.36. The correlation coefficient was positive for the former and negative for the latter. Thus, the higher $\Delta_T$, the higher the ratio $\Delta_{T+\tau}/\Delta_T$ would be. This is to be expected, since larger gaps are more difficult to close than smaller ones. More importantly, the more queries are seen regarding the hospitals, the smaller $\Delta_{T+\tau}/\Delta_T$ will be at $T+\tau$. This suggests a strategic behavior by patients when choosing EDs, i.e., the more people query for wait time information, the smaller the eventual gap.

## 3. Simulation analysis of queueing models

In order to gain a better understanding of the empirical results above, we use simulation models to study factors such as patients' sensitivity to delay, load and network symmetry, on the level of synchronization that can be achieved between two hospitals. We also analyze the impact of the announcement method (delay estimator).

There has been a significant volume of work analyzing the JSQ policy, where customers join the shortest queue among several parallel queues upon arrival. Most of the results are established for networks of single server queues in the heavy-traffic asymptotic regime. The main result of relevance to our discussion is that we only need a small fraction of customers to act strategically (choose the shortest queue) to achieve a high level of resource pooling (state-space collapse in the limit) (Reiman 1984, Turner 2000). In this study, we consider a more complex system—two *multi-server* queues operating in parallel, with customers who choose the queue (server pool) by a rational autonomous decision making approach. Customers choose which queue to join based on a choice model that takes into account the delay factor. Specifically, we use a Multinomial Logit Model (MNL) (see Anderson et al. (1996, §2.6)) where the utility for being served in hospital $i$ with reported delay $r_i$ is $u_i(r_i) = \beta_i - \alpha r_i$. $\beta_i$ is a hospital dependent parameter which reflects the differences between hospitals in terms of their service quality, ED size, etc., that may affect the way customers perceive the value of the service. $\alpha$ measures the cost of delay. The probability of choosing hospital 1 is

$$p_1(r_1, r_2) = \frac{\exp(\beta_1 - \alpha r_1)}{\exp(\beta_1 - \alpha r_1) + \exp(\beta_2 - \alpha r_2)},$$

where $r_1$ and $r_2$ are the reported waiting time of hospital 1 and hospital 2, respectively. Similarly, the probability of choosing hospital 2 is

$$p_2(r_1, r_2) = 1 - p_1(r_1, r_2) = \frac{\exp(\beta_2 - \alpha r_2)}{\exp(\beta_1 - \alpha r_1) + \exp(\beta_2 - \alpha r_2)}.$$

We notice that $\alpha$ measures how sensitive customers are to the differences in reported delays. In other words, given two different reported delays, as $\alpha$ increases, a larger the proportion of patients will choose to join the queue with the shorter reported delay.

We extend the previous JSQ literature in three directions:

a) We introduce a choice model to capture the phenomenon that people are relatively insensitive to small differences in waiting times, but are more sensitive to larger gaps. In our model, we assume everyone chooses which queue to join, but begin to act strategically when they see a relatively large difference in the reported waiting times. Accordingly, we analyze how the sensitivity of customers to delay affects synchronization level of the system.

b) As we are only conducting numerical experiments, our model offers more flexibility in terms of allowing multiple servers for each queue and heterogeneity between the two queues. We also gain more insights into the dynamics of finite scale systems.

c) We analyze how different delay announcements affect the synchronization level of the system. This is motivated by the fact that most hospitals report a 4-hour moving average of historical waiting times instead of the current queue length. In this analysis, we distinguish between the effect of estimator accuracy and the estimation delay. Each factor has a different influence.

In what follows, we shall start with an *idealistic* model where each queue is able to report its true waiting time. Note that if, in addition, $\alpha = \infty$, then *every* arriving customer chooses to join the queue with the shortest waiting time and we achieve complete resource pooling, i.e. the two queues act as a fully pooled one. We analyze how the sensitivity parameter $\alpha$, the offered load (offered load of each system when $\alpha = 0$) and other factors affect the synchronization level and system performance. We then investigate the effect of other waiting time announcements that differ in their accuracy. Lastly, we check how the sensitivity parameter affects the gaps in waiting times between two queues.

The simulation model allows full flexibility in terms of inter-arrival time and service time distributions. For the simplicity of demonstration, we consider the classical Markovian setting where the arrivals follow a Poisson process with rate $\lambda$. The service times are exponentially distributed with rate $\mu$. Without loss of generality, we set $\mu = 1$, i.e. the mean service time as one unit of time. In the ED setting, this one unit of time would be around 4 hours. We denote $W_i(t)$ as the true waiting time of queue $i$ at time $t$, $R_i(t)$ as the reported delay of queue $i$ at time $t$, and $n_i$ as the number of servers in queue $i$. We measure the system synchronization level by the correlation between waiting times as seen by arriving customers.

### 3.1. Idealistic model

In the idealistic model, each queue reports the true waiting time. We distinguish between two cases: symmetric and non-symmetric. In the symmetric case, we assume that both hospitals (queues) have

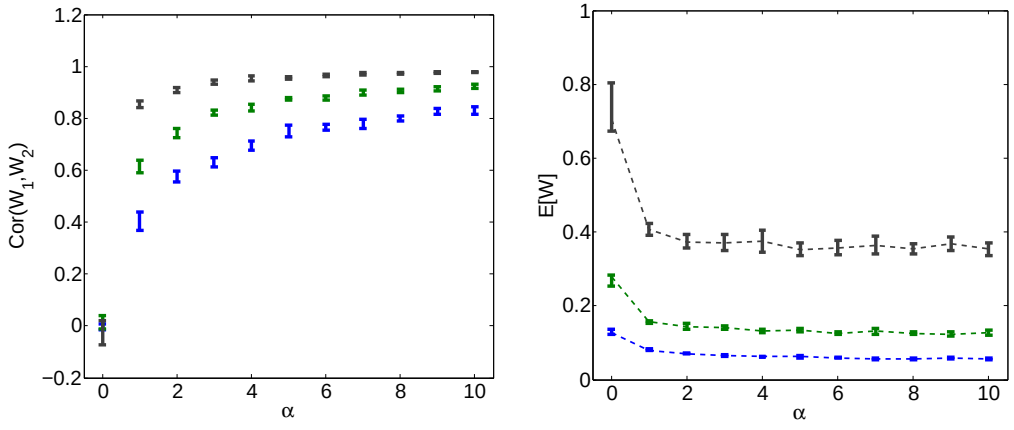
the same number of beds (resources/servers), service time distribution, and are of the same quality. In the non-symmetric case we consider hospitals with different preference parameters ($\beta_1 \neq \beta_2$) and different sizes ($n_1 \neq n_2$).

**3.1.1. Symmetric case** In this case, we assume $\beta_1 = \beta_2$ and $n_1 = n_2$. Hence,

$$p_1(r_1, r_2) = \frac{1}{1 + \exp(-\alpha(r_2 - r_1))}.$$

We analyze how the sensitivity parameter $\alpha$ and the total offered load ($\lambda/((n_1 + n_2)\mu)$) affect the synchronization level between the two hospitals and the system performance—measured by expected waiting times.

For finite values of $\alpha$, our numerical experiments indicate that the synchronization level increases in $\alpha$ (see Figure 7(a)). As previous research on JSQ policy implies, the increase in the sensitivity level is not linear in $\alpha$. We observe that a small value of $\alpha$ will lead to a significant level of synchronization. This suggests that if customers are sensitive only to large gaps between waiting times, we still gain most of the benefit of resource pooling. In addition, we observe that for the same sensitivity parameter $\alpha$, the synchronization level increases with the system load. Figure 7(b) demonstrates that the expected waiting times of both hospitals decreases with $\alpha$. This impact is also influenced by load. As system load increases the effect on expected waiting times is larger, but most of it is achieved even for small values of $\alpha$.




(a) $Cor(W_1, W_2)$

(b) $EW$ (confidence interval: $EW_1$, dotted line: $EW_2$)

**Figure 7** **Synchronization level and expected waiting time as a function of $\alpha$ (upper: $\rho = 0.95$, middle: $\rho = 0.9$, lower: $\rho = 0.85$).**

**3.1.2. Non-symmetric case** Here we consider two non-symmetric cases: quality differences manifested by different $\beta_i$'s and size differences derived by changes in $n_i$'s. The announcements themselves are still exact. We start by assuming that hospitals are of the same size but patients have a predetermined preference towards the second hospital (i.e. $\beta_2 > \beta_1$). This may be due to differences in quality-of-care provided by each institution or some constraints inflicted by the insurance company that drive a higher proportion of the nearby population to a specific facility. As a result, when $\alpha = 0$, $p_1(r_1, r_2) = exp(-\beta_1)/(\exp(-\beta_1) + \exp(-\beta_2)) < 0.5$. Hence, hospital 2 has a higher offer load than hospital 1 to start with, i.e. if we denote $\rho_1 = \lambda p_1(r_1, r_2)/(n_1\mu)$ and $\rho_2 = \lambda p_2(r_1, r_2)/(n_2\mu)$, then $\rho_2 > \rho_1$. We also notice that, since

$$p_1(r_1, r_2) = \frac{\exp(1 - \alpha r_1)}{\exp(1 - \alpha r_1) + \exp((\beta_2 - \beta_1) - \alpha r_1)},$$

we can measure the level of heterogeneity in preference by $|\beta_2 - \beta_1|$.

Observation 3. We make the following three observations from our numerical experiments:

1. Similar to the symmetric case, the synchronization level increases as $\alpha$ increases, and most of the pooling effect is achieved with small values of $\alpha$.

2. For the same value of $\alpha$, the synchronization level is decreasing in $|\beta_2 - \beta_1|$.

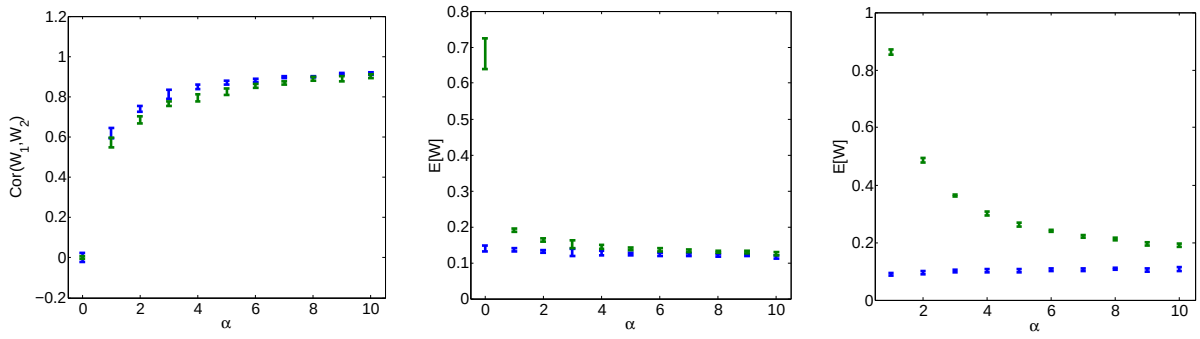
3. The expected waiting time of hospital 2 (the more loaded queue) is decreasing in $\alpha$. The expected waiting time of hospital 1, may decrease or increase in $\alpha$ depending on the preference difference—$|\beta_2 - \beta_1|$. For small value of $|\beta_2 - \beta_1|$, $E[W_1]$ is decreasing in $\alpha$, for large value of $|\beta_2 - \beta_1|$, it is increasing in $\alpha$.

Figure 8 illustrates a numerical example in this setting. We define System A to have $\beta_1 = 1$ and $\beta_2 = 1.1$. In this case, $\rho_1 \approx 0.85$ and $\rho_2 \approx 0.95$. Both $EW_1$ and $EW_2$ are decreasing in $\alpha$. In System B, $\beta_1 = 1$ and $\beta_2 = 2$. In this case, when $\alpha = 0$, $\rho_2 > 1$ (i.e. hospital 2 is unstable when acting alone). Here synchronization assures stability. Nevertheless, this comes at the following cost: $EW_1$ is increasing in $\alpha$ for System B.

We make similar observations when $\beta_1 = \beta_2$ and look at the heterogeneity in staffing levels $(n_1, n_2)$. The synchronization level is increasing in $\alpha$ and the expected waiting time of the more loaded system (system with a smaller staffing level) is decreasing in $\alpha$. For a fixed value of $\alpha$, the synchronization level is decreasing in $|n_2 - n_1|$. For small values of $|n_2 - n_1|$, the expected waiting time of the less loaded system (the system with the larger staffing level) is decreasing in $\alpha$, while for large values of $|n_2 - n_1|$, the expected waiting time of the less loaded system is increasing in $\alpha$.

### 3.2. The importance of timely and accurate delay announcements

The wait time estimator used by hospitals will err from time to time. In this section, we show that these errors limit the synchronization levels that can be achieved and that delays of the wait time




(a) $Cor(W_1, W_2)$ (upper (blue): system A, lower (green): system B)

(b) $EW$ for system A (upper (green): $EW_2$, lower (blue): $EW_1$)

(c) $EW$ for system B (upper (green): $EW_2$, lower (blue): $EW_1$)

**Figure 8 Synchronization level and expected waiting time as a function of $\alpha$.**

estimator in reflecting the true waiting time may drive the system to distraction. We show that the system may become more volatile (oscillating), and that synchronization may decrease (instead of increase) with $\alpha$. Specifically, such behavior occurs when one uses the moving average estimators employed in all the hospitals we observed.

We compare the following two delay estimators:

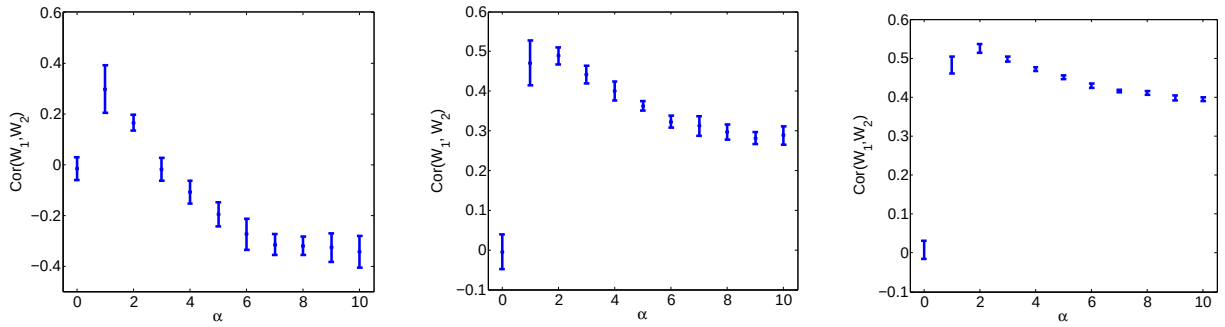
1. *Moving average:* Historical average over time windows of specific lengths. This is the current reporting policy of hospitals. Let $l$ be the time window for the moving average function. We demonstrate two cases: 1. $l$ equals average service time (LOS) (i.e., $l = 1/\mu$); 2. $l$ is one order shorter than the average service time (i.e., $l = 0.1/\mu$).

2. *Head-of-Line (HOL) wait*: Waiting time of the customer that is waiting at the head of the line upon the current arrival. This method could be considered as a using moving average with $l = 0$.

The hospitals in our empirical study report waiting times that are calculated using the average waiting time of patients who enter service during the past 4 hours (moving average of a 4-hour window). This is more or less the case where the averaging window equals the average service time. There are other websites and online apps that report historical averages of even longer periods of time—up to 1 year (see, for example, the online app "ED Wait Watcher" (Groeger et al. 2014)).

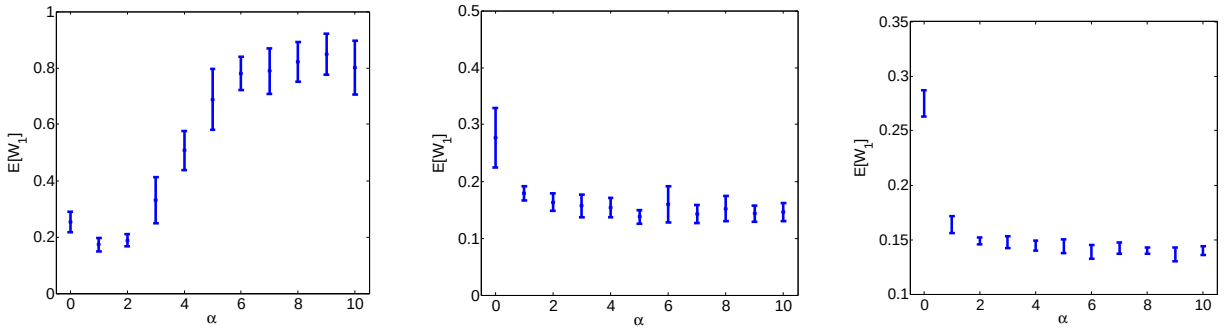
Figure 9 shows how the synchronization level changes with the sensitivity parameter $\alpha$ for different window lengths $l$. Unlike the idealistic model, the synchronization level first increases in $\alpha$ but then decreases in $\alpha$. When $l = 1$, the synchronization level decreases to a *negative* level, while when $l = 0.1$ and HOL ($l = 0$), the synchronization level is positive for all values of $\alpha$. Figure 10 shows how the expected waiting time changes with the sensitivity parameter $\alpha$ for different window length $l$. We observe that when the averaging window is long ($l = 1$), the expected waiting time is *increasing* in $\alpha$ for large values of $\alpha$. This suggests that the system is better off without

announcements if patients are very sensitive to wait. But if one uses a relatively small window ($l < 0.1$) than the expected waiting time decreases as sensitivity increases.




(a) $l = 1$ (b) $l = 0.1$ (c) HOL Wait ($l = 0$)

**Figure 9** **Synchronization level as a function of $\alpha$.**

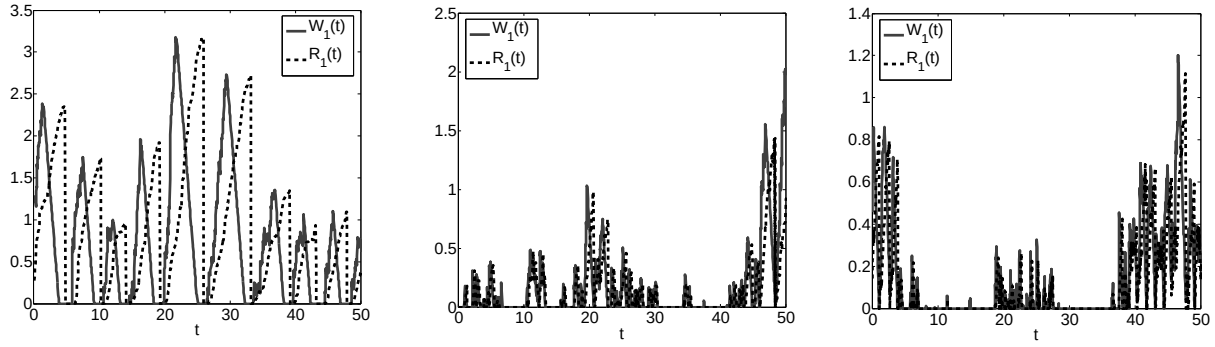



(a) $l = 1$ (b) $l = 0.1$ (c) HOL Wait ($l = 0$)

**Figure 10** **Expected waiting time as a function of $\alpha$.**

Part of this effect is due to differences in accuracy of the estimators. Indeed, the error of the reported waiting times with respect to the true waiting times is increasing in $l$. Specifically, when $\alpha = 10$, $l = 1$, the mean square error of the reported waiting times of hospital 1, $\sqrt{E[(R_1 - W_1)^2]}$, is approximately 1.0901; when $l = 0.1$, $\sqrt{E[(R_1 - W_1)^2]} \approx 0.3269$, and when $l = 0$, $\sqrt{E[(R_1 - W_1)^2]} \approx 0.1906$. Figure 11 shows the sample path of the true waiting times versus the reported waiting times.

Interestingly, the error alone is not what drives the phenomena of performance deterioration with $\alpha$. To validate this, we analyze a *modified-idealistic* model where we report the true waiting time plus an error term that is normally distributed with mean 0 and standard deviation $\sigma$. Large values of $\sigma$ lead to inaccurate delay announcements, but there is no delay effect as with the moving




(a) $l = 1$ (b) $l = 0.1$ (c) HOL wait

**Figure 11** **Sample path of the true waiting time and the report waiting time ($\alpha = 10$) in Hospital 1.**

average method. Figure 12 shows how $\sigma$ affects the synchronization levels. We observe that the synchronization level is monotonically increasing in $\alpha$ for each value of $\sigma$, which is in contrast to the non-monotonic effect of $\alpha$ we observed in Figure 9. Still, the inaccuracy has its impact: synchronization will not reach its full potential when the announcement is inaccurate. This is reasonable, as in that case patient may choose the more loaded queue just because the information was not correct. Hence, as the error increases the maximal synchronization level will decrease.

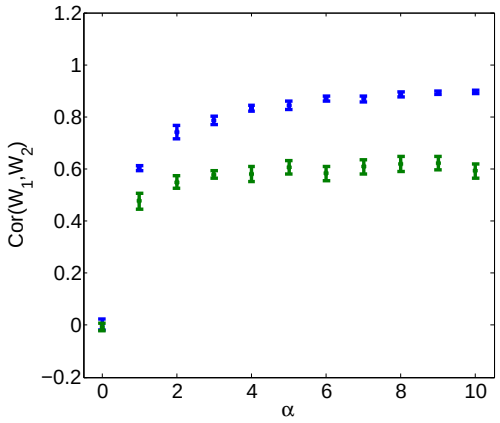
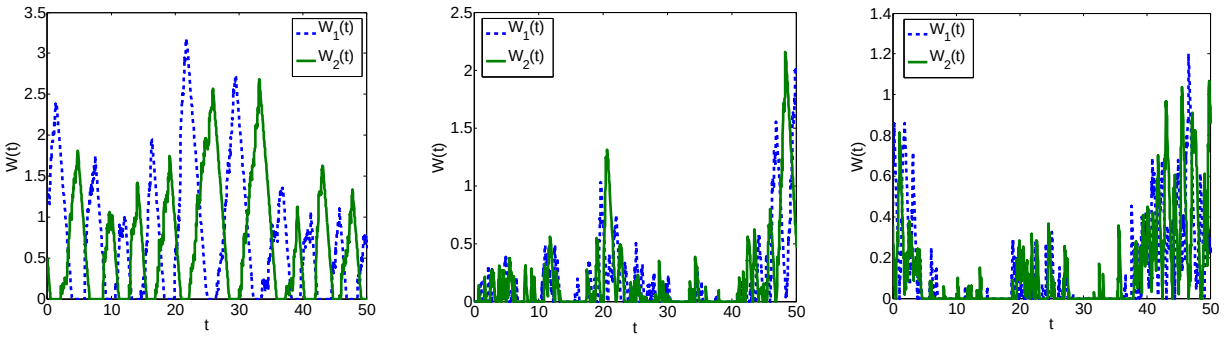



**Figure 12** **Synchronization level for different values of $\alpha$ and $\sigma$ (upper: $\sigma = 0.1$, lower: $\sigma = 1$).**

To understand why a moving average estimator preforms so poorly, we next take a closer look at the sample path of the waiting time process of the two queues for $\alpha = 10$ (see Figure 13). We observe that when $l = 1$, the waiting time processes of the two queues take an alternating oscillating form. Specifically, when $W_1(t)$ is large (small), $W_2(t)$ is small (large). This explains the negative correlation we observe between the two waiting times. When $l = 0.1$ or 0 (HOL), the two waiting time processes are closer to each other. This observation also explains the counter-intuitive result shown in Figure 6, where initially small differences in wait times translated to large differences,

and vice versa. This phenomenon is known in control theory as *self-oscillation*, where systems with a delayed feedback may oscillate solely because of the feedback (Jenkins 2013). Here the delay announcement and its influence on customer choice can be considered a control mechanism. This suggests that the delay effect is the main reason for the "desynchronization" when patients are very sensitive to delay.




(a) $l = 1$ (b) $l = 0.1$ (c) Exact, $\alpha = 2$

**Figure 13 Sample path of the waiting time process ($\alpha = 10$) of the two hospitals.**

### 3.3. Closing the gap in delay times

Inspired by the empirical results of Section 2.2.2, we investigate here how the sensitivity parameter $\alpha$ affects the waiting time gap between two hospitals and the rate in which such a gap will close.

Following the delay announcement analysis we conducted in Section 3.2, we first test how the gaps in delay times, $E[|W_1 - W_2|]$ change with sensitivity parameter $\alpha$ for different delay announcement estimators (Figure 14). We observe that for $l = 1$ (long historical average window), the average gap size is first decreasing and then increasing in $\alpha$, and for short historical average window, the average gap size is decreasing in $\alpha$. We also observe the variation of gap, $Var(|W_1 - W_2|)$, is increasing with the window size $l$. These are consistent with our analysis on delay announcement estimators. We observe that the if the $l = 1$ the gap can be of the same order of the service time, while for more accurate and timely methods ($l < 0.1$) the gap is one order less than service time.)

We next look at how fast a specific gap in waiting time closes for different values of the sensitivity parameter $\alpha$. Specifically, we start the two queues from different initial states, one empty and the other with certain amount of people waiting and observe the duration required, on average, for the two systems to close the gap in waiting times to less than 0.01 units of time, for different values of the sensitivity parameter $\alpha$. We assume the system reports a delay that is based on historical averages of a moving window that is of the same size as the mean service time (to be consistent with our empirical systems). Figure 15 shows the numerical results from simulation. We observe that as

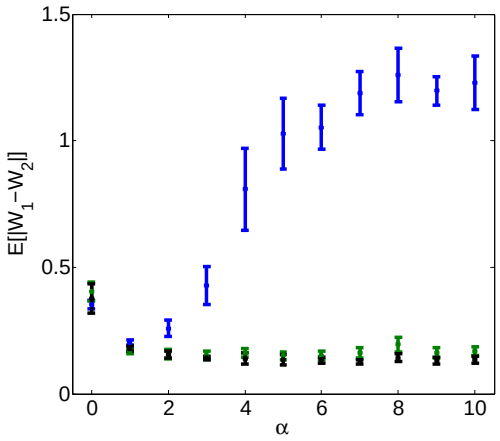



**Figure 14** **Average waiting time gap as a function of $\alpha$ (upper: $l = 1$, middle: $l = 0.1$, lower: HOL ($l = 0$)).**

patients become more sensitive to waiting, more of them choose the hospital with shorter reported (and very likely true) waiting times, thus helping to balance the load faster. We also observe that the higher the total offered load of the system, the *longer* it takes to close such waiting time gaps.

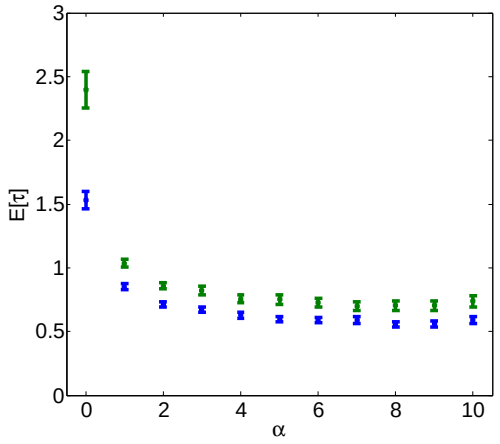



**Figure 15** **Synchronization speed as a function of $\alpha$ (upper: $\rho = 0.95$, lower: $\rho = 0.9$).**

## 4. Conclusion and future research

In this paper, we investigated the impact ED delay announcements have on patients' choices and the effect of patients' choices on hospital synchronization levels. We provide evidence that patients take delay information into account and that they act on this information when the gap between waiting times is large enough. Using numerical simulations we observed that the synchronization level between systems increase with patients' sensitivity to waiting, the load of the system, and the accuracy of the delay announcements. We also found that the mismatch between

delay announcement and actual delay may cause extra oscillations in system load when the patients are very sensitive to delay.

We believe this paper opens several directions for future research. The instability caused by the mismatch between historical average and future delay calls for more accurate machinery for delay announcements. We hypothesize that it is better to provide customers an estimation of the future waiting time, as they need to travel to the ED, and hence do not enter the system immediately. Developing estimators that jointly model travel and wait time is an open question. In addition, the delay until seeing a physician is very partial information of the load in the ED. Another important piece of information is the total LOS in the ED. Estimating future LOS is hard, as the reason that the patient is arriving at the hospital is unknown, and the treatment required for him, as well as the requirements of resources in the system are unknown. Nevertheless, we believe that LOS is an important indicator for patients when considering which hospital to attend. How will providing other load indicators (such as LOS) influence patient choices? Will the fact that they cannot be accurate reduce coordination in the same way? Another issue arises when considering that hospitals can 'play' with the load indicators by changing priorities between new and in-process patients. When giving priority to new patient LOS may increase and vise versa. Does the current system design incentivize hospitals to act in different operational modes as load changes? These are topics we are currently investigating.

## References

Allon, Gad, Achal Bassamboo, Itai Gurvich. 2011. we will be right with you: Managing customer expectations with vague promises and cheap talk. *Operations Research* **59**(6) 1382–1394. doi:10.1287/opre.1110.0976.

Anderson, S. P., A. de Palma, J.-F. Thisee. 1996. *Discrete Choice Theory of Product Differentiation*. MIT Press, Cambridge, MA.

Armony, Mor, Constantinos Maglaras. 2004. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research* **52**(2) 271–292.

Armony, Mor, Nahum Shimkin, Ward Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57**(1) 66–81.

Carmon, Ziv, Daniel Kahneman. 1996. The experienced utility of queuing: real time affect and retrospective evaluations of simulated queues. Working paper, Duke University.

Chalfin, Donald B., Stephen Trzeciak, Antonios Likourezos, Brigitte M. Baumann, R.P. Dellinger. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* **35**(6) 1477–1485.

Duda, Richard O., Peter E. Hart, David G. Stork. 2001. *Pattern classification*. John Wiley and Sons, Inc, New-York, USA.

Foley, Robert D., David R. McDonald. 2001. Join the shortest queue: Stability and exact asymptotics. *The Annals of Applied Probability* **11**(3) 569–607.

Groeger, Lena, Mike Tigas, Sisi Wei. 2014. Ed wait watcher. URL http://projects.propublica.org/emergency/.

Ibrahim, Rouba, Mor Armony, Achal Bassamboo. 2013. Does the past predict the future? the case of delay announcements in service systems. Working Paper.

Ibrahim, Rouba, Ward Whitt. 2009. Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management* **11**(3) 397–415.

Ibrahim, Rouba, Ward Whitt. 2011. 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.

Jenkins, Alejandro. 2013. Self-oscillation. *Physics Reports* **525**(2) 167–222.

Kostami, Vasiliki, Amy R. Ward. 2009. managing service systems with an offline waiting option and customer abandonment. *Operations Research* **11**(4) 644–656.

Larson, Richard C. 1987. OR forum—perspectives on queues: Social justice and the psychology of queueing. *Operations Research* **35**(6) 895–905.

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

Marco, Catherine A., Mark Weiner, Sharon L. Ream, Dan Lumbrezer, Djuro Karanovic. 2012. Access to care among emergency department patients. *Emergency Medicine Journal* **29**(1) 28–31.

Munichor, Nira, Anat Rafaeli. 2007. Numbers or apologies? customer reactions to telephone waiting time fillers. *Journal of Applied Psychology* **92**(2) 511–518.

Plambeck, Erica, Mohsen Bayati, Erjie Ang, Sara Kwasnick, Mike Aratow. 2014. Forecasting emergency department wait times. Working paper, Stanford University.

Reiman, Martin I. 1984. Open queueing networks in heavy traffic. *Mathematics of Operations Research* **9**(3) 441–458.

Senderovich, Arik, Matthias Weidlich, Avigdor Gal, Avishai Mandelbaum. 2014. Queue mining–predicting delays in service processes. *Advanced Information Systems Engineering*. Springer, 42–57.

trends.google.com. 2011–2015. Google trends. URL trends.google.com.

Turner, Stephen R. E. 2000. A join the shorter queue model in heavy traffic. *Journal of Applied Probability* **37**(1) 212–223.

Yu, Qiuping, Gad Allon, Achal Bassamboo. 2014. How do delay announcements shape customer behavior? an empirical study. Forthcoming in Management Science.