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

Jing Dong
Northwestern University, jing.dong@northwestern.edu

Elad Yom-Tov
Microsoft Research, eladyt@microsoft.com

Galit B. Yom-Tov
Technion—Israel Institute of Technology, gality@tx.technion.ac.il

We investigate the impact of delay announcements on the coordination within hospital networks using a combination of empirical observations and numerical experiments. We offer empirical evidence which suggests 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 hospitals use to estimate delays. We show that delay estimators that are based on historical averages 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 occur in hospital networks in the US.

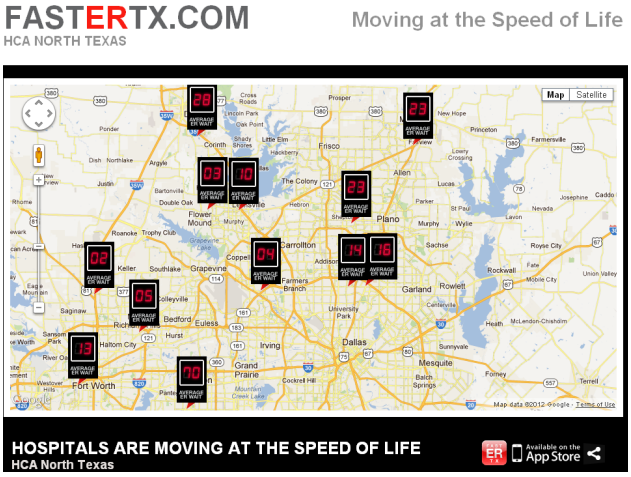
Key words: Delay Announcements, Emergency Department, Queueing Network Coordination, Join the Shortest Queue, load balancing, Cost of Waiting

1. Introduction

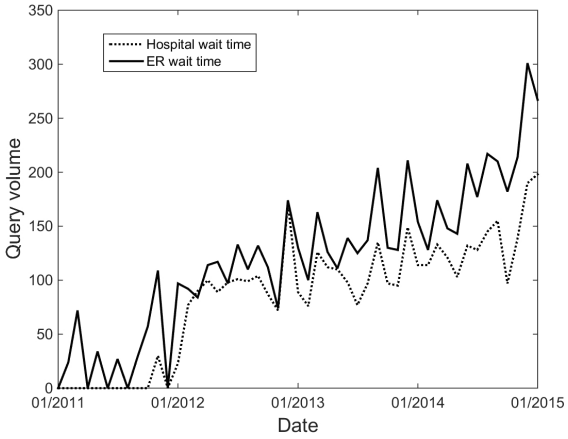
Delay announcements, commonplace in service systems, can be used to influence quality perceptions and customer sentiment towards the service provider. In addition, such announcements can affect customer choices, with follow-on effects on actual system operations. In consequence, delay announcements have over recent years attracted the attention of the operations research and management communities, with research streams dedicated to both understanding the impact of delay announcements and developing methods to support them. Thus far, most research in this area has concentrated on call center announcements, where delay information has been shown to influence 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(a))—a trend that evidence suggests is welcomed by consumers. As can be seen in Figure 1(b), the volume

of Google search engine queries (as reported in trends.google.com) for "hospital wait time" and "ER [Emergency Room] wait time" has been rising steadily over the past several years. Yet it is unclear whether and how such information actually affects patients' choices—and the subsequent effects of their choices on hospitals' performance. Although patients' primary consideration in selecting an ED is generally its timely provision of treatment, other factors have bearing as well, including the reputation of the hospital, its expertise, limitations imposed by patients' medical insurance plans, and recommendations by the primary physician (Marco et al. 2012). Given the effort required to provide waiting time information for hospital ED services, a number of questions should be addressed. First, do customers actually want this information and do they use it? Second, is the proportion of people who seek such information large enough to have an operational impact on the healthcare system, and on hospital networks in particular? Third, do hospitals provide the right information to help achieve coordination (load balancing) in the network?



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



(b) Google Trends query volume regarding ED wait times

Figure 1 ED wait time Internet query trends and real-time delay announcements

Here we examine the effect of ED delay information (i.e., delay announcements or waiting time announcements) on patients' choice of hospital emergency departments, using a combination of empirical analysis and numerical experiments. We base our analysis on two primary sources of data: real ED delay announcements from more than 200 hospitals in the US over a period of three months; and anonymized queries made to the Bing search engine by people seeking ED delay information during that period. The empirical data provide objective evidence for the use of delay information, including public interest in such delay announcements and the influence of such announcements on the delays themselves.

Drawing on insights from queueing theory, we then study the operational influence of delay announcements on correlations among different hospitals' waiting times. We refer to the correlation between the wait times of two hospitals (or EDs) as *synchronization*, and to EDs whose wait times are positively correlated as synchronized EDs. We then use a stylized simulation model, calibrated with real data, to investigate how system characteristics, such as patients' sensitivity to waiting, load, and different delay estimators, influence the phenomena observed in the data.

1.1. Scientific Background

Delay announcements have a measurable influence on customer satisfaction (Carmon and Kahneman 1996, Larson 1987). However, such announcements can vary widely in their specificity, from vague information on the current load to relatively precise details about the customer's location in the queue or their expected waiting time. The effects of these messages differ. Munichor and Rafaeli (2007) showed that, in a call center environment, informing customers about their location in the queue results in lower abandonment rates and higher customer satisfaction compared to other waiting time fillers, such as music or apologies. Allon et al. (2011) developed a game theoretic model, based on a strategic service provider and a strategic customer, which provides a theoretical basis for determining how vague or specific delay announcements should be.

One of the challenges in implementing 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, as well as time-varying, arrival process (Ibrahim and Whitt 2009, 2011). Their proposed estimators are based on a real-time history of the queue, including last-to-enter-service (LES) information (i.e., the delay experienced by the last customer entering service) and head-of-line (HOL) data (the total delay experienced by the customer currently at the head of the line). 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 theory-based queueing predictors with real data.

Estimating delays in EDs is substantially 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 undergo a process involving multiple resources (physicians, nurses, labs, etc.) which generally take 3 to 6 hours to complete. The complexity is even greater when we consider that patients' arrival rate is time-varying, patients are prioritized according to severity, and that any given patient's route is unknown ahead of time. Plambeck et al. (2014) developed a forecasting method for estimating ED delays based on a combination of queueing and machine learning methods. None of the hospitals from which we drew our data use such sophisticated models. Instead,

they publish historic average waiting times using a 4-hour moving average, a measure which has become the convention in US hospitals.

As suggested above with respect to call centers, delay announcements influence not only customer satisfaction but also customer waiting costs and, in response, customer actions. Announcing the expected delay as customers enter the system, especially during heavily loaded periods, may cause customers to balk (leave the system upon arrival) or abandon after a short time (Mandelbaum and Zeltyn 2013, Yu et al. 2014). Xu et al. (2016) show that information provided to potential patients on availability and quality of service influences physician demand. In this paper, we use a multinomial choice model to incorporate the effect of a delay announcement on customer's arrival decision (as in Armony et al. (2009)). A similar choice model has been applied in Huang et al. (2013) to measure the effect of bounded rationality, that is, the ability of customers to estimate delay.

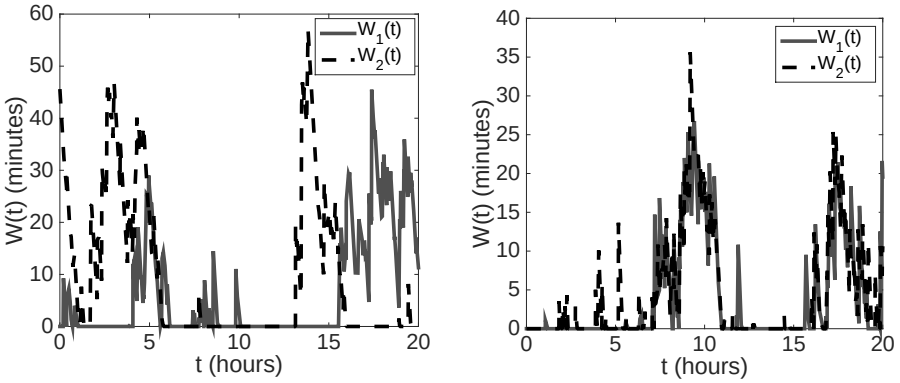
Given the potential influence of delay announcements on customer behavior, the announcements can be used as an operational tool. For example, delay announcements are used in call centers to help customers choose their time of service via a call-back option (Armony and Maglaras 2004). In theme parks, delay announcements can enhance resource allocation by helping customers choose preferred queues (Kostami and Ward 2009).

The above-mentioned research investigated the impact of delay announcements on the company which provides the information. Our paper investigates the impact of such announcements in a network setting, as is the case when several EDs are located in the same area. In such settings, announcements by one service provider may impact demand at other providers. Moreover, in the case of EDs, service providers are not only in competition, but also have incentives to cooperate. On the one hand, hospitals want to attract patients, and EDs are considered the 'gateway' to a range of hospital services. On the other hand, the expensive nature of ED services limits their resources and capacity—leading to occasional high congestion and long waiting times, with the potential for diminished quality of care (Chalfin et al. 2007) and increased mortality (Bennidor and Israelit 2015). At such times, therefore, the hospital has an incentive to ease some of the load by reducing the arrival rate.

Some hospitals attempt to share the load through ambulance diversions. However, ambulance diversions are inefficient for this purpose because the patients being diverted in this way are those most in need of urgent medical care. Therefore, some US states ban this policy (e.g. Massachusetts). The goal of hospitals is, instead, to influence the behavior of those least in need of immediate treatment: the non-acute patient population, which can account for up to 90% of ED visits (Plambeck et al. 2014). Delay information, unlike ambulance diversions, can be used when the ED is crowded to encourage those non-acute patients to respond to long wait times by choosing a different ED or

delaying their visit. Assuming patients take such information into account, announcements are thus expected to smooth the demand for hospital services throughout the day and to balance patient loads between nearby hospitals.

Figure 2 compares the simulated wait-time sample path of two networks, each with two hospitals. Figure 2a shows the wait times of two hospitals, where patients uniformly randomly choose which hospital to attend. We observe that wait times are uncorrelated. Figure 2b shows the same simulation when patients always choose the hospital with the shortest waiting time. We base our analysis on the observation that in the latter, the wait times of the two hospitals are synchronized (i.e., the wait times of the hospitals are highly correlated). Therefore, in this paper, we identify connections between delay announcements and correlations between workloads at geographically proximate EDs. We also observe that the waiting time of the second system is much shorter than (about half) the waiting time of the first system, on average. This is because synchronization leads to load balancing, and thus better performance. If wait time can be halved in reality by the mere announcement of wait time, it has the potential to improve patient care and safety dramatically with very low costs.



(a) Two independent (uncorrelated) queues

(b) Two fully synchronized (correlated) queues

Figure 2 Unsynchronized vs. synchronized systems.

From a theoretical point of view, even a small amount of coordination—for example, a system that enables a fraction of customers to choose to Join the Shortest Queue (JSQ)—improves the efficiency of the network to almost the level of a fully pooled system (Foley and McDonald 2001, Turner 2000). Hence, one of the operational rationales for providing delay announcements is that they may improve the efficiency of the hospital network.

This paper uses online search engine logs to explore correlations between waiting time observations and people's choice of emergency departments. Search engine queries have been shown to

reflect people's activities in the physical world as well as the virtual one. For example, Ofran et al. (2012) found a high correlation between the number of searches for specific types of cancer and their population incidence. Similarly, a high correlation was observed between the number of searches for certain medications and the number of prescriptions written for those drugs (Yom-Tov and Gabrilovich 2013). Additionally, search engine logs have been used to monitor influenza activity (Polgreen et al. 2008), to examine the association between online exposure to underweight celebrities and the development of eating disorders (Yom-Tov and boyd 2014), and to discover potential adverse effects of certain medicines (Yom-Tov and Gabrilovich 2013). In the context of emergency departments, aggregated search logs were used to predict ED visits (Ekström et al. 2014).

1.2. Goals and main contributions

The main questions we address in this work are: Is there empirical evidence that patients are influenced by delay announcements when choosing an emergency service provider? If so, how do patients' choice and system characteristics influence waiting time synchronization within the hospital network?

Combining empirical analysis and numerical experiments, we make the following contributions:

• We develop a new performance measure to define the level of load balancing a queueing network.

• Using that measurement and by utilizing search engine data we find that:

— Patients explore delay information provided by hospitals, and do so at a growing rate.

— There is significant variation in the level of load balancing each hospital network experiences, and about 15% of those networks experience negative pairwise correlation (§3.1.2).

— The number of customers who take delay information into account when selecting ED providers is sufficient to have a load balancing effect within the hospital network. We show that by exploring periods in which hospitals randomly stopped providing data, and show that providing such information has a significant effect on wait times of adjacent hospitals (§3.2.2).

— The distance between hospitals is also correlated with synchronization: the greater the distance between two hospitals, the lower their effect on one another (§3.1.3).

— Increasing the number of hospitals reporting wait time (in an area) increases synchronization, but the number of potential customers who are exposed to wait time information has a non-linear effect on synchronization levels in hospital networks (§3.2.3), which we further explore in the numerical section of the paper.

• Synchronization between two hospitals is influenced by how sensitive customers are to delay, the load of the system, the scale (size) of the system, and the hospitals' heterogeneity in terms of customer preferences and size (§5).

- If the appropriate method is used for delay estimation (and loads are fairly balanced), the social welfare of patients in this network increases. That is, delay announcements will reduce average wait times for *all* hospitals in the network.
- The accuracy of the wait time estimator, as well as the delay of the estimator in reflecting true wait times, 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 high oscillation of the load between hospitals and increase wait times when there is a large enough proportion of strategic patients and/or when patients are very sensitive to delay (§5.2).

2. Data description and transformation

2.1. Data description

To analyze patient choices, we use information on the announcements provided to potential visitors, as an indication that potential patients were exposed to that knowledge, and data on the outcome of such choices on the network state. We augment the data with control variables to explain heterogeneity in population between cities throughout the US.

Hospital level information - Real-time announcement data: For the first data source, we identified 211 US hospitals which published their waiting times using RSS feeds as of March 2013. RSS (Really Simple Syndication) is a method Internet websites can use to provide updates to online content in a standardized, computer-readable format (Libby 1999). Note that, as explained below, publication of wait times on an RSS feed does not entail publishing this information in human-readable format on the hospital website.

We collected these waiting times every 5 minutes by polling the hospitals' RSS feeds between April and June 2013 (inclusive). The waiting times refer to the time expected to elapse from when the patient enters the ED to when he/she is first seen by a qualified medical professional[1]. No real-time information is provided on the patient's total length of stay or classification according to severity. Nevertheless, all the hospitals explain in their websites that the reported wait times do not apply to urgent patients, as these patients are prioritized according to their medical condition. Hence, hospitals urge patients to ignore this information if they are in immediate threat to their lives. All the hospitals included in this research use the same estimation method, namely, a moving average over a 4-hour time window. From our data it appears that this information is updated every 15 minutes. Note that the higher sampling rate utilized here (once every 5 minutes) is required because the sampling is not synchronized to the hospital's rate of change, and thus the higher rate reduces the lags in the time it takes to identify a change in reported wait times.

[1] A qualified medical professional is defined as a Doctor of Medicine (MD), Doctor of Osteopathy (DO), Physician Assistant (PA) or Advanced Registered Nurse Practitioner (ARNP).

All the hospitals in our sample published their wait times in their RSS feed. However, not all hospitals published the wait time information on their websites for human consumption in the same manner. Some provided only their own waiting times, some also provided information on waiting times of (two or more) nearby hospitals, and some did not present their wait times on the hospital website at all during those 3 months. In all hospitals, if wait time information was presented, it was presented on the front page of the hospital website. In our data, 27 hospitals did not present any wait times, 104 showed only their own wait times, and 80 showed their wait times as well as those of nearby hospitals.

Geographic and demographic information: The second data set includes data about the area in which each hospital operates. These include the number of other medical facilities in the area, distances between hospitals, as well as demographic information on the population living in the area of each hospital.

To estimate the total number of hospitals in each area we obtained a list of 36,438 medical facilities listed in the Bing Local search application, and their location. As this list contains irrelevant medical facilities (maternity centers, mental health facilities, and rehabilitation facilities), we filtered the list to include only those hospitals registered with Medicare[2], resulting in a total of 4,576 hospitals throughout the US. This information allows us to compute the distances between hospitals.

As our analysis focuses on how patients choose between geographically proximate hospitals, the 211 hospitals that report wait times, in the RSS feeds, were divided into clusters such that the reporting hospitals in each cluster were within no more than 20 km apart from one another (see §3.2.3). This partition created 46 clusters of hospitals, each defining a geographical area enclosing the reporting hospitals within it.

For each such cluster, the following additional variables were calculated:

- Fraction of hospitals reporting their own wait times;
- Fraction of hospitals reporting their own wait times as well as those of proximate hospitals;
- Fraction of hospitals reporting their own wait times and requiring another click to present those of proximate hospitals.

Several demographic variables may influence both access to online information as well as the tendency to use it (Perrin and Duggan 2015). Hence, we collected demographic county-level information from three sources:

1. The US Census American FactFinder[3].

[2] `https://data.medicare.gov/Hospital-Compare/Hospital-General-Information/xubh-q36`

[3] `http://factfinder.census.gov/`

2. The US Census Small Area Income and Poverty Estimates[4].

3. US Government Health Indicators Warehouse[5]

The demographic variables includes:

- Number of primary care persons per 100k population;
- Poverty level;
- Household income (HHI);
- Median male age;
- Median female age;
- Male to female ratio;
- Fraction of children aged 4 or under;
- Fraction of people 65 and over;
- Total population.

Remark: We note in passing that wealth does not seem to be associated with more hospitals publishing their wait times. To demonstrate this, we extracted the adjusted gross income (AGI) from the 2012 IRS data for the closest zip code to each of the 4,576 hospitals identified in the Bing local search application (as explained above). The median AGI matched to hospitals that published wait times is not dissimilar from that of all other hospitals ($P = 0.03$, ranksum test).

Exposure information: Finally, to measure interest and exposure to the presented information, we extracted all queries made using the Bing search engine between April and June 2013 (inclusive) which resulted in a visit to the web page of one or more hospitals on our list. We assume that all visitors to these pages noticed the wait times, which are prominently displayed.

Each query contained an anonymized user identifier, a time stamp, user location details (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 the query.

Following Yom-Tov et al. (2013), we define sessions as all contiguous queries made by a user without a break of 30 minutes or more. We note that 2.2% of sessions where users inquired about the location of a hospital or its emergency room resulted in visits to the websites of more than one of the target hospitals. This indicates that a non-negligible percentage of users seek information about multiple hospitals when choosing which one to visit.

Bing is the second-largest search engine operator in the USA, with an estimated market share of approximately 20%, as of March 2015[6]. We assume that the population of Bing users is a good

[4] https://www.census.gov/did/www/saipe/data/statecounty/data/2013.html

[5] http://www.healthindicators.gov/Indicators/Primary-care-providers-per-100000_25

[6] http://www.comscore.com/esl/Insights/Market-Rankings/comScore-Releases-March-2015-US-Desktop-Search-Engine-Rankir

representation of the general population. This assumption is supported by the following analyses: First, the correlation between the number of Bing users per county in the USA and the number of people in that county according to the 2010 US Census is $R^2 = 0.83$ ($P < 10^{-6}$). Second, an analysis of data collected from an opt-in consumer panel recruited by an Internet analytics company comScore, which includes age (in 5-year increments) and gender, shows a correlation of $R^2 = 0.62$ ($P = 0.004$) of the fraction of users between Bing and Google at each age and gender group.

From the exposure data above we calculated the following variables:

- Number of queries made to the Bing search engine about each of the reporting hospitals;
- Number of Bing queries about each hospital which also mentioned the word 'wait';
- Number of Bing queries per capita.

In addition, for each area of clustered hospitals we defined the following variables:

- Number of Bing sessions which included queries about more than one hospital, where one was to a hospital in the cluster;
- Number of EDs reporting wait times per square km within a cluster;
- Number of hospitals per square km (whether reporting wait times or not) within a cluster;
- Number of pediatric EDs within the cluster (some EDs do not cater to children, or do so only during certain hours).

2.2. Data transformation

The main explanatory variables in our analysis are hospital waiting times and their correlation among nearby hospitals. However, these wait times are reported in a way that requires pre-processing to remove the effects of recurrent patterns and reporting biases caused by averaging.

2.2.1. Removing diurnal and weekly patterns The wait times of hospitals have strong daily (diurnal) and weekly patterns of activity, exhibiting higher wait time during the day than during the night. Hence, closely situated hospitals will have a natural synchronization level that is due to the pattern of the exogenous arrival rate, and not necessarily to endogenous influence of information provided to potential patients. In most of the empirical analyses, we would like to exclude such synchronization.

To focus on correlation due to endogenous influence of information provided to potential patients, we remove the hourly and daily waiting time trends from the observed (reported) waiting times, for each hospital, in the following manner. For each hospital, a linear predictor was trained to predict waiting times from the hour of the day and the day of the week ($h(t) \in [0, 1, ..., 23]$ and $d(t) \in [1, 2, ...7]$). These predicted waiting times were subtracted from the observed waiting times. Thus, we denote the waiting time predicted from the hour and day as r_i^{Rec}, since it captures the

recurrent portion of the reported wait time signal, and refer to the resulting detrended reported waiting times as Residual Waiting Times (RWT).

Let $RWT_i(t)$ be the residual wait time of hospital i at time t. Then,

$$r_i(t) = \beta_1 h(t) + \beta_2 d(t) + \epsilon_i(t);$$
$$r_i^{Rec}(t) = \beta_1 h(t) + \beta_2 d(t);$$
$$RWT_i(t) = r_i(t) - r_i^{Rec}(t).$$

The average ratio between the weight of the hourly trend and the weight of the daily trend was 0.999, indicating that the terms had almost identical effects. Together, these two variables explain 85.0% of the variance of the reported waiting times, i.e. the trained $r_i^{Rec}(t)$ achieves an $R^2 = 0.85$.

2.2.2. The effect of averaging on wait time correlations The wait time data collected represents precisely the information provided to potential visitors, but is not an accurate indicator of the *actual* hospital state. This is because wait times are provided after averaging using a 4-hour moving window[7], and rounding to the nearest minute.

If the exact (e.g., not rounded) wait time would have been provided, the non-averaged wait time could be reconstructed up to a constant. However, attempting to reconstruct the non-averaged wait time from the available signal quickly diverges from the average wait times, as the rounding errors accumulate and can be larger that the actual wait times.

However, below we show that an approximation can be elicited for the cross-correlation between the reported hospital wait times.

Let $w_i(t)$ be the *actual* wait time of hospital i at time t. Without loss of generality let $w_i(t)$ be normalized to zero mean and unit variance. The effect of averaging using a moving window is akin to convolving $w_i(t)$ with a rectangular window W of length 4 such that:

$$f(t) = \begin{cases} 1/4 & \text{if } 0 \le t < 4; \\ 0 & \text{otherwise.} \end{cases}$$

The *observed/reported* averaged wait time, $r_i(t)$, is thus denoted by $r_i(t) = w_i(t) * f(t)$, where $*$ denotes the convolution operator.

The cross-correlation between two signals $x(t)$ and $y(t)$ can be computed as $x(t) * y(t)$. For two wait times, $w_i(t)$ and $w_j(t)$, after averaging, the correlation of the observed signal is:

$$C^{Obs} = (w_i(t) * f(t)) * (w_j(t) * f(t)). \quad (1)$$

[7] For example, the Medical Center of Lewisville, one of the reporting hospitals, states on its website that "ER wait times represent a four-hour rolling average... defined as the time of patient arrival until the time the patient is greeted by a qualified medical professional."

Using properties of the convolution operator:

$$C^{Obs} = (w_i * w_j) * (f * f). \tag{2}$$

Thus, the correlation of the averaged wait times is equal to the correlation of the unaveraged wait times, convolved with a window which equals:

$$F[t] = f * f = \begin{cases} (4+t)/16 & \text{if } -3 \le t < 0; \\ (4-t)/16 & \text{if } 0 \le t < 4; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

This is a triangular window centered at 0. The effect of this convolution is that the observed correlation from the (averaged) wait times is a weighted average of the non-averaged correlation at lags of between -3 and +3 hours.

The relationship between C_{ij}, the (unfiltered) cross-correlation function of x_i and x_j, and the observed cross-correlation C_{ij}^{Obs} for a signal of length N can be written explicitly (recalling that $F[-i] = F[i]$) in matrix form as:

$$\begin{pmatrix} F[0] & F[1] & ... & F[4] & 0 & 0 & 0 & ... & 0 \\ F[1] & F[0] & F[1] & ... & F[4] & 0 & 0 & ... & 0 \\ F[2] & F[1] & F[0] & F[1] & ... & F[4] & 0 & ... & 0 \\ \vdots & & & & & & & & \\ 0 & 0 & \cdots & 0 & F[4] & F[3] & F[2] & ... & F[4] \end{pmatrix} \cdot \begin{pmatrix} C[0] \\ C[1] \\ C[2] \\ \vdots \\ C[N] \end{pmatrix} = \begin{pmatrix} C^{Obs}[0] \\ C^{Obs}[1] \\ C^{Obs}[2] \\ \vdots \\ C^{Obs}[N] \end{pmatrix} \tag{4}$$

In matrix form, we denote this as $F^M \cdot C = C^{Obs}$. Since F^M is a fully-ranked (symmetric) Toeplitz matrix, C can be recovered by the Moore-Penrose pseudo-inverse:

$$C = (F^M \cdot F^M)^{-1} \cdot F^M \cdot C^{Obs} \tag{5}$$

This correction is true up to the rounding discussed above. Rounding errors, which are akin to quantization errors, can be modeled as additive white noise (Widrow et al. 1996). This, together with the fact that the cross-correlation function averages over a large number of samples, should cause rounding to have a negligible effect, especially for shorter lags.

2.2.3. Correcting detrended wait time correlation for smoothing effects As explained in 2.2.1 we would like to measure the correlation of the residual wait time. In this section we show that these correlations can be estimated either by first correcting for smoothing and then removing the recurring trends, or vice versa. In practice, the latter is preferred, since it is less sensitive to rounding effects in the data.

As noted above, the wait time of a hospital can be decomposed into two components: A recurring diurnal and weekly component, and a component representing all other effects. These two

components are provided to us after smoothing and rounding, so that the observed wait time $r_i(t)$ can be written as:

$$r_i(t) = w_i * f = (w_i^{Tr}(t) + w_i^{Rec}(t)) * f \tag{6}$$

where $w_i^{Rec}(t)$ is the recurring component, and $w_i^{Tr}(t)$ the component representing all other transient effects of the real wait times.

Without loss of generality, if both $w_i^{Rec}(t)$ and $w_i^{Tr}(t)$ are time series with zero mean and unit variance, the correlation between two hospitals before correction for smoothing can be written as:

$$E\left((w_i^{Tr}(t) * f)(w_j^{Tr}(t) * f)\right) = E\left(\left((w_i(t) - w_i^{Rec}(t)) * f\right)\left((w_j(t) - w_j^{Rec}(t)) * f\right)\right) = \\ E\left(\left(w_i(t) * f - w_i^{Rec}(t) * f\right)\left(w_j(t) * f - w_j^{Rec}(t) * f\right)\right) \tag{7}$$

The components $w_i(t) * f$ and $w_j(t) * f$ are simply the observed time series, while $w_i^{Rec}(t) * f$ and $w_j^{Rec}(t) * f$ are the recurring components, convolved by the smoothing operator. When the recurring components are estimated after smoothing (as described in Section 2.2.1 above), they correspond to these smoothed recurring components.

Therefore, to accurately estimate the correlation between the detrended time series of two hospitals we can simply calculate the correlation between the detrended time series (Section 2.2.1), and then correct it using the procedure in Section 2.2.2. We denote the detrended and corrected correlation between hospital i and j by C_{ij}^{TR}. In the remaining of the paper, unless otherwise stated, correlations between pairs of hospitals are corrected in this manner.

3. Empirical Results

In this section, we draw on the data described in Section 2 for objective indicators that patients indeed use delay information in choosing emergency service providers, and that this has a profound influence on network coordination.

We begin our discussion with a descriptive analysis of the variations in waiting times, the effects of time-of-day, and the difference between the announced waiting time to the actual waiting time and its influence on our analysis. We then report our findings on the effects of wait time reporting on hospital synchronization.

3.1. Initial investigation

3.1.1. Wait time variation Hospitals differ greatly in their reported wait times. To examine these variations, we clustered the 211 hospitals which reported their wait times into three groups according to their wait times, using the K-means algorithm with Euclidean distance. Each hospital was represented by a vector of the average hourly reported wait times.

The resulting clusters partition the hospitals into three groups: low wait, medium wait, and high wait times. Table 1 provides characteristics of each cluster. Figure 3 shows the average of the reported wait times, $r_i(t)$, for the three groups over a 3-week period. The figure demonstrates daily patterns as well as the variation in waiting times among the three groups. The average wait times in highly loaded hospitals range from 30 minutes at night to 70 minutes wait during the day. Moderately loaded hospitals experience wait times ranging from 10 to 30 minutes, while lightly loaded hospitals announce average wait times of 7 minutes during the night to 12 during the day.

Cluster	Low wait	Medium wait	High wait
Number of hospitals	97	99	13
Average wait (min)	8.2	16.4	32.5
Standard deviation (min)	3.1	2.7	8.2

Table 1 Clustering hospitals by reported wait. N=211

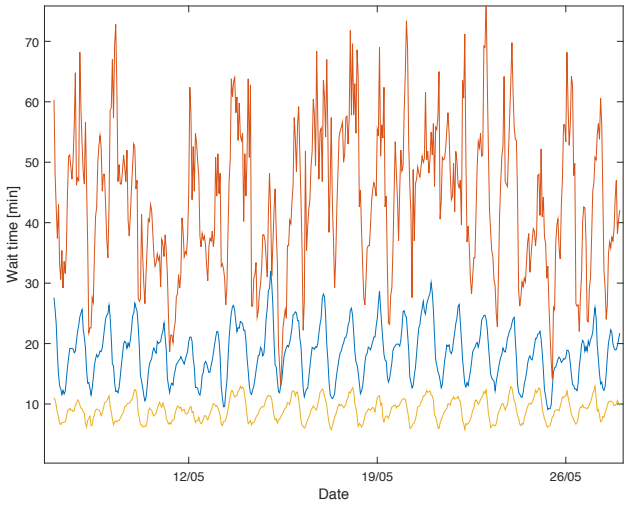


Figure 3 Average waiting times for the three groups of hospitals—low wait, medium wait, and high wait times.

3.1.2. Correlations between hospital wait times Figure 4 shows a histogram of the Spearman correlations found between wait times of all pairs of hospitals. The line denoted as 'reported' represents correlation of the original wait time signals as related to visitors to the website, and the line marked as 'detrended' shows the correlation of the RWTs.

We observe that, as expected, the correlation of the reported wait times is greater than the correlation of detrended wait times. In general, most hospital pairs show positive correlations, but surprisingly, a significant proportion of them show negative correlation. We will discuss possible reasons for such phenomena in Section 5.2.

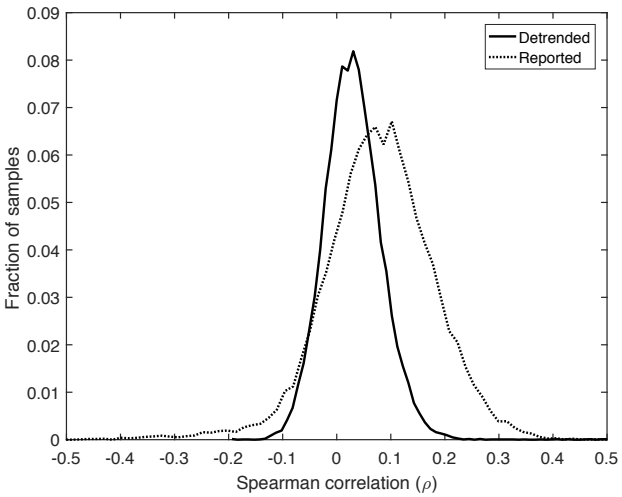


Figure 4 **Correlation of reported waiting times versus correlation of detrended waiting times.**

3.1.3. The effect of distance between hospitals on synchronization Figure 5 plots the wait times of two hospital pairs. As the figure suggests, some ED pairs are much more synchronized than others, a fact which is also evident in Figure 4. An important factor that is correlated with synchronization level is distance.

Recall that $C_{i,j}^{Obs}$ is the correlation between the observed wait time of hospital i and hospital j, e.g., $r_i(t)$ and $r_j(t)$, respectively, over the 3-month period. Let $DIST_{i,j}$ be the distance between hospital i and hospital j. The correlation between the distance among hospitals, $DIST_{i,j}$, and the level of synchronization between them, $C_{i,j}^{Obs}$, is negative (Spearman, $\rho = -0.138$ ($P < 10^{-5}$)).

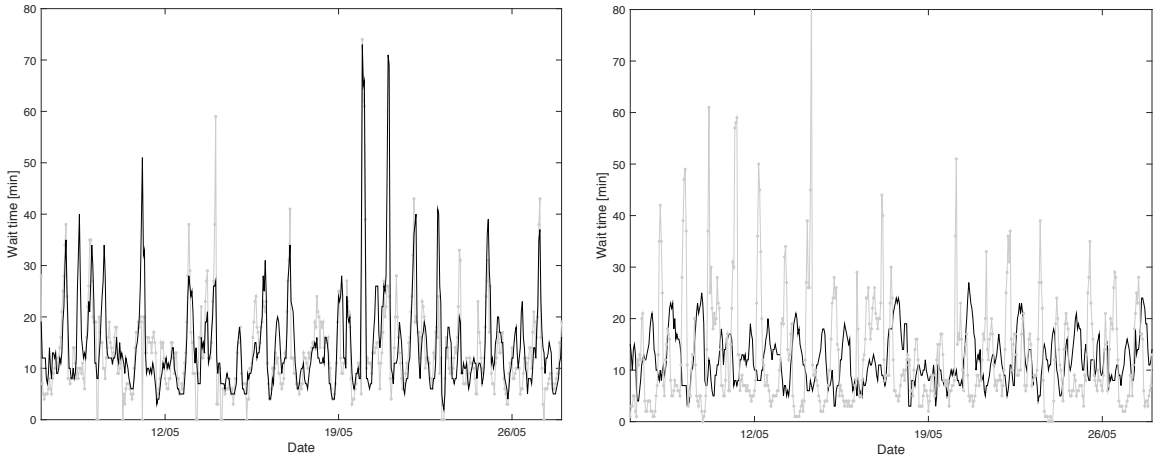
Observation 1. Distance relates to synchronization; greater distance between pairs of hospitals is associated with a lower correlation between their wait times.

This geographical synchronization could be attributed to several factors, including, daily patterns and information provided to ambulance services. What we seek next is to understand whether *announcements* of anticipated delays contribute to this synchronization, and to what extent.

3.2. The association between delay announcement and wait time synchronization

3.2.1. How does the amount of information displayed correlate with hospital synchronization? Hospitals display wait time information in various ways. While some hospitals suppress the display of that information, all the hospitals in our dataset which present that information do so on their front page. Still differences in display occur; some hospitals show only their own ED information, while others show the information of nearby hospitals as well. We identify in our dataset four levels of information display:

- Not presented: Wait times are not shown on the hospital website.
- Only self: Only the own hospital's wait times are shown.



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

Figure 5 Waiting times announced by different EDs.

- Others too: Both the hospital's own wait times, as well as that of nearby hospitals are shown.
- Others, click: The hospital's own wait time is shown, and that of adjacent hospitals are shown when the user clicks on a button to display them.

Figure 6 presents examples of the three display formats.

(a) Only self (b) Others too (c) Others, click

Figure 6 Examples of a hospital wait time displays.

Figure 7 shows the average correlation of the RWT, C^{TR}, according to the *level* of information displayed in the hospital website, stratified by the distance to the nearest hospital reporting wait times. The figure suggests that nearby hospitals that display their own wait times as well as that of nearby hospitals are more correlated than those that do not. This correlation is much lower when distance to the nearest hospital is large (grey bars in Figure 7). These observations indicate that the information display coupled with the availability of close alternatives are associated with an increased correlation in wait times.

We evaluate this hypothesis using an ANOVA model, where the dependent variable is the correlations among hospitals, and the independent variables are the information display type, the distance (continuous), and the interaction of the two. The results are presented in Table 2, showing

Figure 7 **Average RWT correlation (C^{TR}) by information display and distance to nearest hospital. Correlations when the nearest hospital is less than 8km apart are shown in strip bars, and among farther hospitals in gray. Correlation of wait times differ significantly by both distance and information type (See Table 2.**

that both variables and their interaction are statistically significantly correlated with the correlations among hospitals. Note that there is no strong statistical evidence about which type of display (only self, others, click, or yes) is most effective. In what follows, we distinguish between whether the hospital displays the wait time information or not.

Variable	F	p-value
Information type	3.32	0.024
Distance between hospitals	6.01	0.016
Interaction between the two variables	2.79	0.044

Table 2 **ANOVA model for the correlation between nearby hospitals as a function of information type displayed and the distance between hospitals. The model is constructed for hospitals 30km or less of each other.**

3.2.2. The effect of withholding information on wait times In this section, we evaluate whether providing information to potential patients through a website creates a load balancing effect and therefore reduces wait times. Sometimes, probably due to technical faults, the wait times disappear from the RSS feeds. These can be used to construct "natural experiments" (Meyer 1995) to establish causality between the effect of one hospital reporting wait times and the wait times of the closest hospital to it (neighboring hospital). In particular, when one hospital ceases publishing its wait times, we cannot measure synchronization. Instead, we can measure the effect that the disappearance of the wait times in one hospital has on the wait time of its neighboring hospital. Since the wait time information is not provided, load balancing reduces or stops. Our conjecture is that wait times in the neighboring hospitals will increase during such breaks. However, the effect of stopped load balancing should only appear in hospitals where such information was provided to

patients. If such information was not provided both before and after the stop, no significant change should be observed. Therefore, we hypothesize the following:

HYPOTHESIS 1. ED wait times are shorter when wait time information is displayed in two adjacent EDs as opposed to when information is not published in both hospitals.

Figure 8 demonstrates the change in RWT of the closest hospital (but not more than 20 km distant) to the one in which the reporting break has happened, from 12 hours before the break in wait time displays and until 12 hours after it. In order to set a similar baseline for all hospitals, we removed the average RWT in the 12 hours preceding the break from the entire (25-point) data sequence for each break, and show the median times in Figure 8. As the figure shows, there were smaller changes in the wait times of hospitals which did not report wait times. However, those that did report it saw a large increase in their wait times. This is the expected trend: when information stops being presented, load balancing stops too, and therefore waiting times increase.

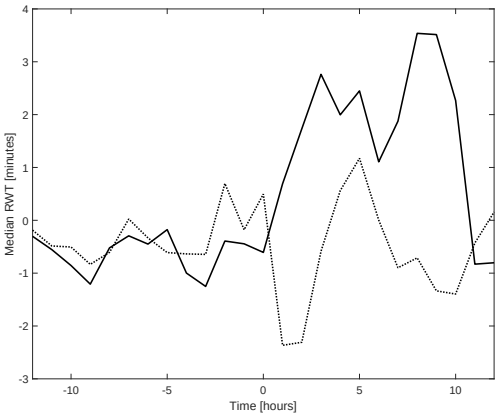


Figure 8 **Median RWT of the closest hospitals when wait time reports disappear. The horizontal axis denotes time, where zero hour is when the break in wait time reporting began. The dotted line shows the median RWT of the closest hospital, when the break occurred in a hospital which did not display wait times on its website ($n = 45$), while the full line shows the same when the break occurred in those hospitals that showed the wait times for both them and adjacent hospitals ($n = 25$).**

As a more rigorous statistical test, we perform a difference-in-difference analysis of the breaks. We focus on times where wait times were absent from the RSS feed for 4 hours or more, stratified by the way in which they were normally displayed in the hospital website. We found in our data 2740 such 4-hour breaks; 1771 of those occurred in hospitals that did not display the wait time information on their websites. One may suggest that hospitals stop providing information strategically, for example, when they are extremely loaded and prefer not to present information which may show

them in unfavorable light. To refute this, we checked that the breaks do not appear to be correlated to wait times. Specifically, the wait times reported one hour prior to the break, compared to the average of the wait time in the corresponding hour one day before and one week before at the same hospital are not statistically significantly different (signtest, $P > 0.05$). Additionally, the day of the week at which breaks occur is not statistically significantly different (chi^2 goodness-of-fit test, $P = 0.92$). Interestingly, breaks do tend to concentrate during a particular time period of the day, with the average at 6AM and the mode at 3AM. Thus, we deduce that long breaks occur mostly due to technical failure, when technical staff are not available to fix the failure.

To test Hypothesis 1, we construct the following difference-in-difference model (Abadie 2005) at the hospital level:

$$\bar{T}_{i,j,t} = \beta_0 + \beta_1 \cdot h_i + \beta_2 \cdot I_t + \beta_3 \cdot Disp_i + \beta_4 \cdot Disp_j + \beta_5 \cdot DIST_{i,j}\beta_6 \cdot (I_t \circ Disp_i) + \beta_7 \cdot (I_t \circ Disp_j) + \beta_8 \cdot (I_t \circ Disp_i \circ Disp_j) + \epsilon_{i,j,t} \quad (8)$$

Here, $\bar{T}_{i,j,t}$ is the average wait time at hospital j, which is the closest hospital to hospital i, in the time interval t (either 4 hours prior to the break or the first 4 hours following the break); $DIST_{i,j}$ is the distance between the two hospitals; h_i is the time of day in which hospital i experienced the break; I_t indicates whether t was a break interval; $Disp_i$ indicates whether hospital i displays wait times on its website in general; $Disp_j$ indicates whether hospital j displays wait times on its website in general; $(I_t \circ Disp_i)$ is an interaction term equal to 1 if t is a break interval and hospital i displays information in general; $(I_t \circ Disp_j)$ is an interaction term equal to 1 if t is a break interval and hospital j displays information in general; $(I_t \circ Disp_i \circ Disp_j)$ is an interaction term equal to 1 if both hospitals i and j display information and t refers to time during a break; β_i's are the corresponding coefficients; and $\epsilon_{i,j,t}$ is the remaining error term.

In estimating Equation 8, we use the estimator of the interactions to compare the difference in hospitals' average wait time before break occurred to the difference after the break occurred. Because the information to which the potential patients were exposed was not different before and after the break if hospital i does not display information, we consider hospitals that do not provide information ($Disp_i = 0$) as the untreated comparison group and those hospitals that display information ($Disp_i = 1$) as the treatment group. The use of the difference-in-differences approach enables us to control the effects of variables that are common to both hospitals before and after the breaks, even when those variables are unobserved.

There should also be a difference for whether hospital j displays information or not. If hospital j does not display wait time information on its website, patients who are considering going there will not be able to compare wait times regardless of the information presented by hospital i.

Therefore, the group for which we expect to see a difference when breaks occur is one in which both hospitals display information, i.e. both $Disp_i = 1$ and $Disp_j = 1$. Therefore, we estimate the effect of publishing wait time information by examining the coefficient on the interaction term, β_8.

Table 3 presents the coefficient estimated for Equation 8. We observe that if a hospital displays wait times, it reduces the wait times of the adjunct hospital ($\beta_3 = -0.51$, $p = 0.0004$), but once that information stops we see an increase in wait times only if both hospitals display information ($\beta_8 = 1.11$, but $\beta_6 = -1.20$).

Number	Variable	Parameter estimate (SE)	p-value
1	h - Time of day	0.07 (0.02)	0.0004
2	I_t - During break?	-1.71 (0.56)	0.0022
3	$Disp_i$ - Hospital i reports wait times	-0.51 (0.14)	0.0004
4	$Disp_j$ - Hospital j reports wait times	0.83 (0.51)	0.1011
5	Distance between hospital i and j	0.47 (0.16)	0.0046
6	Interaction of I_t and $Disp_i$	-1.20 (0.52)	0.0221
7	Interaction of I_t and $Disp_j$	1.31 (0.56)	0.0195
8	Interaction of I_t, $Disp_i$ and $Disp_j$	1.11 (0.52)	0.0352

Table 3 Model of the effect of withholding information.

This natural experiment provides supporting evidence that the information provided in the hospital's website indeed influences patient choices, and thus influences the synchronization levels by creating load balancing in the network. We next provide a more detailed analysis of factors that are associated with the level of synchronization.

3.2.3. How does network structure and exposure influence load balancing? As noted above, the past several years have seen a steady growth in the number of people searching online for delay information. In the following we investigate how load balancing effects are influenced by the network structure in which hospitals operate and people's tendency to look for delay information online. To do this, we model the level of synchronization between clusters of geographically proximate hospitals (less than 20 km apart) using geographic, demographic, and exposure information as explanatory variables.

To define 'geographically proximate hospitals', we clustered the hospitals which published wait times according to their geographic location by performing Agglomerative Hierarchical Clustering (Duda et al. 2001) with closest-link aggregation until no hospitals were found within 20 km. Specifically, each hospital was initially considered a separate cluster. We then iteratively merged the pair of clusters closest to each other (in the sense of the closest-link), until the closest link was greater than 20 km. This procedure resulted in 46 clusters.

We propose two models for these data: A non-linear model and a linear model with interactions between variables.

Regression tree model: For the non-linear model we first trained a regression tree (see Figure 9) to predict the average correlation C^{TR} within a cluster using the geographic, demographic, and exposure variables defined in Section 2.1. Applying the leave-one-out cross-validation method to estimate the performance of the model, we found that the Spearman correlation between predicted and actual C^{TR} values was $\rho = 0.528$ ($P = 0.002$), indicating that the variables provide good predictive power for the dependent variable—i.e., the synchronization level as measured by the correlations between RWT.

The variables of the highest levels of the tree (Figure 9) are the number of EDs reporting wait times per unit area within a cluster, the number of Bing queries about hospitals in the cluster, median female age, and the fraction of children aged 4 or under in the area of the cluster. We attribute the first variable to the availability of wait time information, where more information translates to a higher correlation. The last two attributes are demographic variables, which imply that specific populations are more likely to engage in decisions on which hospitals to visit. For example, young people are more likely to use delay information, which is not surprising since young people are more likely to use the Internet as a source for health information (Fox and Duggan 2013). A surprising observation is that the number of queries has a non-monotonous effect. In some populations and networks, more queries are associated with lower synchronization levels, which is the opposite of what we expected. We will discuss possible explanations for that phenomena in Section 5.2 via controlled simulation experiments.

Stepwise linear regression model: These data were also modeled by a stepwise linear regression model, which is shown in Table 4. Similar to the regression tree, this linear model shows that the number of reporting EDs is the most important factor. However, instead of the total number of queries per cluster, we observe that the number of queries about *multiple* hospitals becomes significant here. A third interesting factor is the relative number of primary care personnel in the area, that is significant only through interaction with the number of reporting EDs. The coefficient of the interaction is negative, suggesting that if there are many EDs in the area but also many primary care physicians, there is less synchronization. Our explanation is that in such areas only more urgent patients seek ED services, and the delay information is less relevant for them. This conjecture is based on the fact that in the US, EDs are used as a supplement of primary care service when the latter is not well provided (NEHI 2010).

Observation 2. Customer's exposure to delay information is associated with hospital synchronization. As is evident from Table 4 and supported by Figure 9, the number of delay-reporting hospitals per unit area and the number of queries about waiting times are significant indicators of hospital network synchronization. The higher the number of reporting hospitals per unit area and the greater the number of queries, the higher the correlations observed. When coupled with

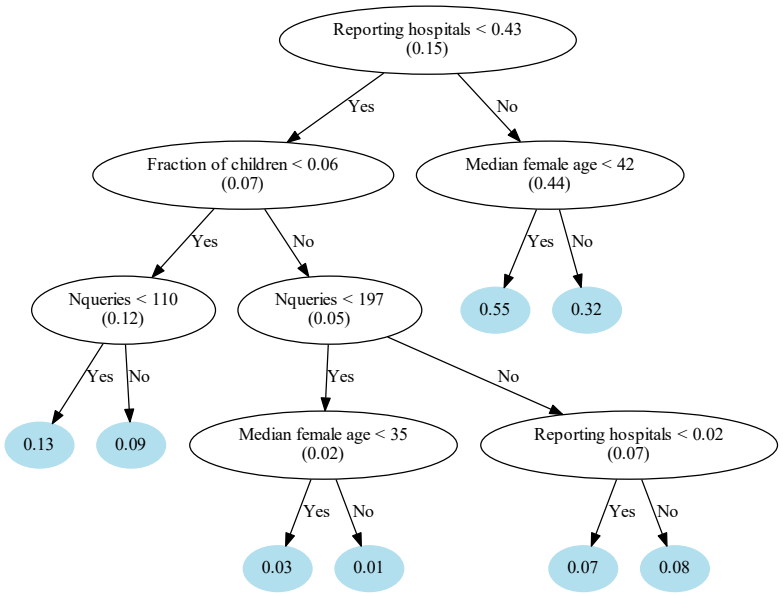


Figure 9 **Regression tree classifier for predicting in-cluster RWT correlations. Decision nodes show the splitting variable and the average correlation at the node (in parentheses). Leaf nodes show the average correlation at the node. The split point is given in each node. For example, "Reporting hospitals $<$ 0.43" means that all clusters where the fraction of reporting hospitals per square km is lower than 0.43 will be routed to the left branch, and all others to the right one. The variables are listed in the main text.**

Variable	Parameter estimate (SE)	p-value
(1) Number of EDs reporting wait times per square km within a cluster	3.90 (0.37)	$< 10^{-10}$
(2) Number of Bing queries about multiple hospitals in the cluster	0.004 (.002)	0.046
(3) Number of primary care persons per 100k population in the cluster area	50.3 (30.2)	0.10
(4) Median male age	0.009 (0.003)	0.003
Interaction of (1) and (3)	-1255.6 (148.5)	$< 10^{-10}$
Interaction of (1) and (4)	-0.051 (0.008)	$< 10^{-6}$

Table 4 **A stepwise regression model for the average correlation within a cluster ($R^2 = 0.90$, $P < 10^{-6}$).**

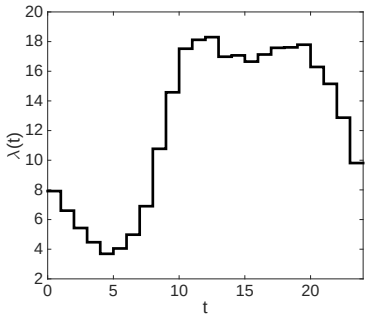
the observation of Section 3.2.2, we conclude that the proportion of customers that take delay information into account is large enough to have an operational influence on the state of EDs and their delay times.

4. A simulation case study

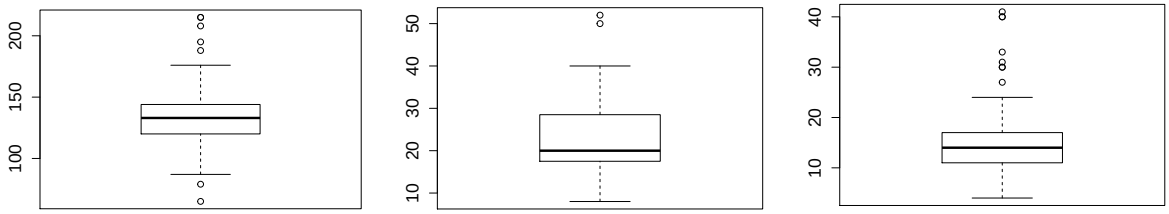
Our empirical study provides evidence that the number of EDs reporting waiting time information and the exposure level to these data are important factors in predicting the synchronization level of hospital waiting times. Our observations also raise questions on specific aspects of synchronization, e.g., the exact role of information seeking (number of queries) and synchronization rates, as well as the effect of different demographics. However, the nature of our data, which does not include specific patient level choice information, limits our ability to establish the causal relationship between

patients' reaction to delay announcements and the synchronization level of waiting times. In this section, we use stochastic simulation to mimic the dynamics of a network with two hospitals in the same neighborhood and investigate how patients' sensitivity to delay may affect the synchronization level of hospital waiting times. We will show, using simulation study that the level of synchronization observed in the data can indeed result from patients acting strategically to delay information. Later, in Section 5, we will study factors that influence the level of synchronization.

Model calibration: In this study, we model the two hospitals as two multi-server queues with a time-varying arrival rate and time-varying staffing levels, $M_t/G/s_t$. The arrivals follow a time-varying Poisson process. We fit the arrival rate function using the scaled arrival rate data of all ED visits in the US during 2010 (Centers for Disease Control and Prevention 2010), which is a piecewise constant function as depicted in Figure 10(a). This shape of arrivals is common in many EDs throughout the world (e.g. Armony et al. (2015), Shi et al. (2015)). The most common Length of Stay (LOS) in EDs follows a Lognormal distribution (see e.g. Armony et al. (2015)). Hence, we assume service times are independent and identically distributed Lognormal random variables with parameter $\mu = 4.18$ and $\sigma = 1$. In particular, the mean service time is 108 minutes, which is the national average LOS of patients who are discharged after ED treatment minus the national average waiting times for those patients. We notice that according to the Medicare website, the average LOS of more severely ill or injured patients, who need to be hospitalized after ED treatment, is 274 minutes, which is much longer than the LOS we use. We choose to focus on the less urgent cases because we think these are the patients who will use delay information to choose which hospital to visit. The limited capacity in the ED is a combination of the number of physicians and the number of beds. Many hospitals in the US apply a case load management policy, where each physician has a cap on the maximum number of patients they can care for simultaneously. Hence even if beds are available in the ED, the capacity might be limited due to the number of physicians attending. A typical capacity constraint is the number of attending physicians multiplied by the number of patients per physician. A description of the connection between the two and its impact on waiting can be found in Song et al. (2015) and Campello et al. (2017). We will refer to this limiting capacity in the general term 'servers'. We use two staffing levels: from 7:00AM to 11:00PM; both hospitals are staffed by 15 servers, and from 11:00PM to 7:00AM the next day, both hospitals are staffed by 10 servers. We notice that hospital EDs in general have three shifts which results in two different staffing levels (Song et al. 2015). Figure 10 (b)–(d) summarizes data on the average LOS of patients who are discharged after ED treatment and the average ED waiting time (both collected from the Medicare website) and the size of the ED as measured by the number of beds (based on data from 24 hospitals which publish this information on their websites).



(a) Arrival rate function (per hour)



(b) Average LOS (minutes) (c) Number of beds (d) Average waiting time (minutes)

Figure 10 Model calibration data summary

Customer choice: Many factors come into play when patients select an ED. These include, for example, the reputation of the hospital, its expertise, and any limitation imposed by medical insurance plans (Marco et al. 2012). Those are in addition to the delay information discussed here. Moreover, not all potential patients are exposed to delay information, and the weight of this factor for those that are exposed to it is unclear. Therefore, we consider a choice model that incorporates several factors that may affect how patients choose between two hospitals.

We call the patients who check the delay information strategic patients. We denote θ as the proportion of the strategic patients among all arriving patients. We assume the non-strategic patients are equally likely to choose any one of the two nearby hospitals. We model the choice of the strategic patients using a Multinomial Logit Model (MNL) (see Anderson et al. (1996, §2.6)). Specifically, the utility for being seen in hospital i with reported delay r_i is

$$u_i(r_i) = \beta_i - \alpha r_i + \epsilon_i,$$

where ϵ_i is an unobservable patient dependent term that is assumed to be i.i.d. Gumbel with parameter 0 and 1, β_i is a hospital-dependent parameter that reflects differences between hospitals in terms of their service quality, insurance policies, etc., which may affect how patients perceive the "value" of the service, and α measures the "cost of delay". We do not assume outside alternatives

and the utility can be negative, i.e. the patient must choose one of the two hospitals. The probability of choosing hospital 1 then takes the form

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

where r_1 and r_2 are the reported waiting times of hospital 1 and hospital 2, respectively. By rearranging the above equation, we have

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

When the difference between the reported waiting times of the two hospitals is small, the strategic patients will choose the less loaded one with slightly greater probability. When the difference between the reported waiting times is large, the strategic patients will almost certainly (with probability 1) choose the less loaded one. For the same waiting time difference, the larger the value of α, the more likely the patient will choose the less loaded one. We refer to α as the sensitivity parameter.

For our simulation study in this section, we set $\beta_1 = \beta_2 = 1$, and test the effect of different values of θ and α on the synchronization level between the resulting wait times of hospital 1 and 2, $Cor(w_1, w_2)$, and on the correlation between the detrended wait times $Cor(w_1^{Tr}, w_2^{Tr})$. The system performances are measured by long-run average waiting times of the two hospitals, $E[w_1]$ and $E[w_2]$. All the correlations and expectations are calculated based on the long-run time average, i.e. $E[w_i] = \lim_{t\to\infty} \frac{1}{t}\int_0^t W_i(s)ds$, and $Cov(w_1, w_2) = \lim_{t\to\infty} \frac{1}{t}\int_0^t (W_1(s) - E[w_1])(W_2(s) - E[w_2])ds$. We assume the existence of these limits, as the arrival rates and staffing levels are periodic.

Reported waiting times (waiting time estimator): The hospitals in our empirical study estimated waiting times using a 4-hour moving average. We use the same moving average as the reported waiting times of the two hospitals.

Simulation calibration results: Figure 11 plots the average hourly wait times. We notice that the time variability of the wait time process agrees with our empirical data of the high loaded hospital group.

Figures 12 & 13 summarize the simulation results of this case study with its corresponding 95% confidence intervals. We start both hospitals from empty. Each simulation run contains 1000 days of data where the true waiting times are recorded every 15 min. We apply a burning period of 100 days. The estimators and the corresponding confidence intervals are constructed using the Batch means method with 10 batches. As the average wait times of the two hospitals are very close to each other (due to symmetry of the system parameters), we only plot the wait times of hospital 1.

We observe that when there is only a small fraction of people acting strategically ($\theta = 0.1$), the synchronization level increases with α (Figure 12a) and the average wait time decreases with α

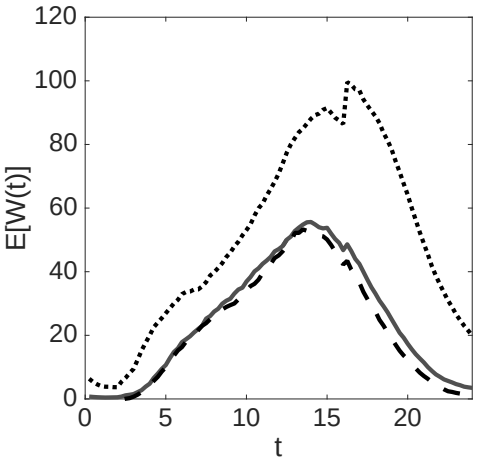


Figure 11 **Expected waiting time for different values of θ ($\alpha = 0.05$, solid: $\theta = 0$, - -:$\theta = 0.1$, $\cdots$:$\theta = 1$)**

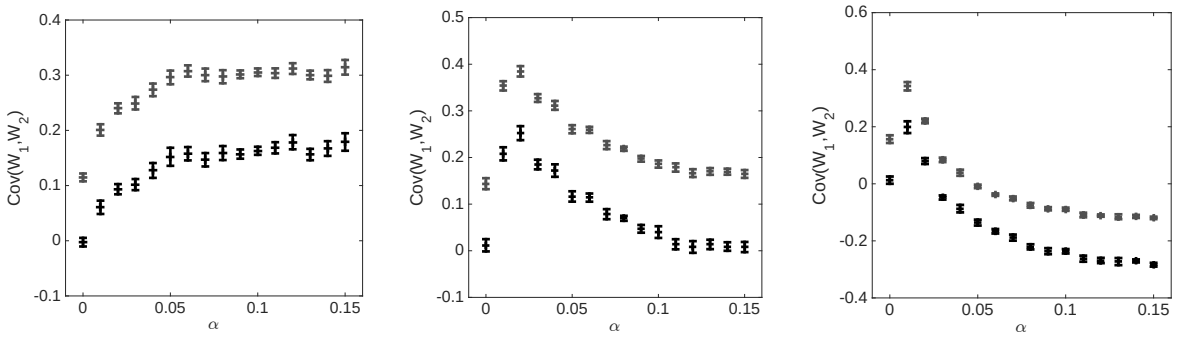
(Figure 13a). These changes in wait times is in line with the changes in the synchronization level, and the classical theory of the effect of the JSQ policy. In particular, the higher the synchronization level, the higher load balancing we can achieve, and the better the performance. The increase in the synchronization level is concave, i.e. it increases faster for small values of α.

Surprisingly, when θ increases things change: When there are more people acting strategically ($\theta = 0.5$), the synchronization level first increases and then decreases with α. When $\theta = 0.9$, the synchronization level can decrease to negative values for large values of α. This phenomenon requires further elaboration, which we provide in Section (§5.2), showing that this is mainly due to the delay effect in the waiting time estimator (four-hour moving average).

We notice, in our simulation case study, that reduction in the expected waiting time is at most around 16%, which is not as much as we would expect from a fully synchronized system (Figure 2). This is because the detrended synchronization level ($Cor(w_1^{Tr}, w_2^{Tr})$) we can achieve is very small, around 0.2 at its maximum. Note that this was also the case in the data we collected, in which the range of synchronization was below 0.2, and could indeed be negative for some hospital groups. Hence, this case study captures some of the observations seen in the real world.

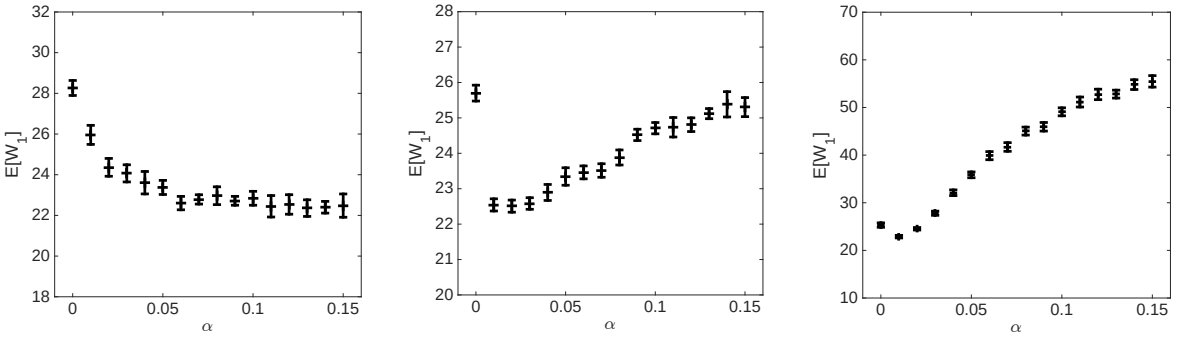
Additionally, we see that θ and α have a similar effect. This precludes our ability to estimate a realistic α. Nevertheless, if one could estimate the proportion of people looking for information, then α could be estimated using our simulation.

In the next section, we will take a closer look at the stochastic model we suggested and investigate what limits our ability to achieve a higher synchronization level.



(a) $\theta = 0.1$ (b) $\theta = 0.5$ (c) $\theta = 0.9$

Figure 12 **Synchronization level between the two hospitals for different customer sensitivity values and the proportion of strategic customers. (Correlation between actual wait times is presented in gray, while correlation between detrended wait times is presented in black.)**



(a) $\theta = 0.1$ (b) $\theta = 0.5$ (c) $\theta = 0.9$

Figure 13 **Expected waiting time of the two hospitals for different values of customer sensitivity values and the proportion of strategic customers.**

5. Sensitivity analysis of the queueing models

To improve our understanding of the level of synchronization that can be achieved with different parameter values, we use simulation models to study the impact of the following factors: load, system scale, network symmetry and different waiting time estimation methods (delay estimator).

As the proportion of strategic customers, θ, and the sensitivity to wait of strategic customers, α, have a similar effect on the synchronization level, in this section, we fix $\theta = 1$ and focus on changes in the patients' sensitivity to delay only. We also assume both the arrival rate and the staffing level to be time homogeneous. This simplification will allow us to directly remove the effect of time variability on the correlation between waiting times. We verified that all results follow if arrivals are time-varying.

The simulation model allows full flexibility in terms of inter-arrival time and service time distributions. For simplicity, we consider the classical Markovian setting where arrivals follow a Poisson process with rate λ. Service times are exponentially distributed with rate μ.

There has been a significant volume of work analyzing the JSQ strategy, where customers join the shortest of several parallel queues upon arrival. Most results in this area are established for networks of single-server queues in the heavy-traffic asymptotic regime. The main insight related to our work is that we only need a small fraction of customers to act strategically (choose the shortest queue) to achieve a high level of load balancing (state-space collapse in the limit) (Reiman 1984, Turner 2000).

Note that we are looking at a system in which customers choose a queue not according to the length of the queue but according to the expected wait in the queue. Hence, the analysis is actually of a Join the Shortest Wait (JSW) policy. Nevertheless, we expect the two policies to preform in a similar manner (Selen et al. 2016).

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. We analyze how the sensitivity of customers to delay (the value of α) affects the synchronization of the system. The choice model also allows the inclusion of preferences which are not related to delays at each ED (by varying β_i), thus allowing heterogeneity between hospitals.

b) We focus on pre-limit performance. We gain more insights into the dynamics of small-scale systems. As we are only conducting numerical experiments, our study offers flexibility in terms of allowing multiple servers in each hospital (queue), heterogeneity in hospital sizes within the network, time-varying arrival rates, and general service time distributions.

c) We investigate the effect of the type of information provided to customers on the synchronization of the system. This is motivated by the fact that most hospitals report a moving average of historical waiting times instead of the true waiting time or the current queue length. We are also able to distinguish between the effect of estimator accuracy and the time lag (delay effect) of different delay estimation methods.

We shall start in Section 5.1 with an *idealistic* model where each system has perfect information and hence 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 load balancing, i.e. the two queues act as a fully pooled one. We analyze how the sensitivity parameter α, the offered load (i.e., the offered load of each system when $\alpha = 0$), system scales, asymmetries in patients' preferences, and system sizes affect synchronization and system performance. Then, in Section 5.2, we investigate the effect of non-perfect information in which waiting time announcements (delay estimator) may be inaccurate.

As before, we denote $w_i(t)$ as the true waiting time (delay) 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 for $i = 1, 2$.

Similarly, the reported simulation results are based on a single long run of 1000 days of data. The waiting times are recorded every 15 minutes. Both systems start from empty and we discard the first 100 days of data as the burning period. We report both the estimator and the corresponding 95% confidence interval using the method of batch means with 10 batches.

5.1. Perfect information (idealistic) model

In the ideal model, each hospital will be able to provide patients with the precise waiting times (this is only possible in a simulation model). As customers' cost of waiting increases, patients are more sensitive to small differences in wait times, and we would expect to see a higher synchronization level. Our focus here is to understand how other system parameters may limit the synchronization level that can be achieved for various sensitivity values.

We divide the analysis into two cases: symmetric and non-symmetric. In the symmetric case, we assume that both hospitals (queues) have the same capacity, service time distributions, and quality (i.e. patient preferences are equal). In the non-symmetric case we consider hospitals with either different preference parameters or different capacities.

5.1.1. Symmetric case: The impact of the cost of waiting (α), offered load and system scale Here we assume that $\beta_1 = \beta_2$, $\mu_1 = \mu_2 = \mu$ and $n_1 = n_2$ and analyze how the total system load, $\rho = \lambda/((n_1 + n_2)\mu)$, and system scale, n_1, affect a) the synchronization between the two hospitals, measured by the correlation between waiting times, and b) system performance, measured by average waiting times.

In this case, the probability of choosing Hospital 1, when reported w_1 and w_2 as the wait times of Hospital 1 and Hospital 2, respectively, is:

$$p_1(w_1, w_2) = \frac{1}{1 + \exp(-\alpha(w_2 - w_1))}.$$

Our numerical experiments lead to the following observations:

OBSERVATION 3. a) Sensitivity to wait times increases synchronization, and greater synchronization leads to better performance for both queues.

b) Load increases synchronization.

c) Smaller-scale systems gain greater synchronization.

We next demonstrate Observation 3 through the following numerical examples. Figure 14 plots synchronization as a function of α for three different values of system load ρ. We observe that synchronization increases with α, and the increase in synchronization is not linear. A small value of α will lead to a significant increase in synchronization. This suggests that even if customers are only

sensitive to very large gaps between waiting times, we still achieve a high degree of synchronization. For fixed α, when $\alpha = 0$, the load has no effect on the synchronization level; when $\alpha > 0$, the more loaded the hospitals, the more synchronization we gain.

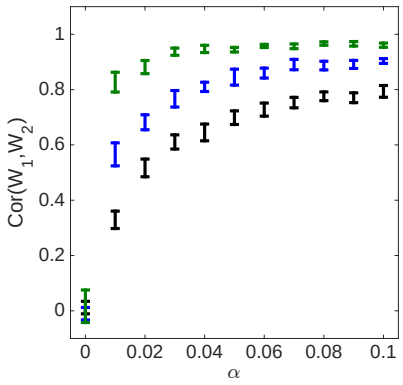
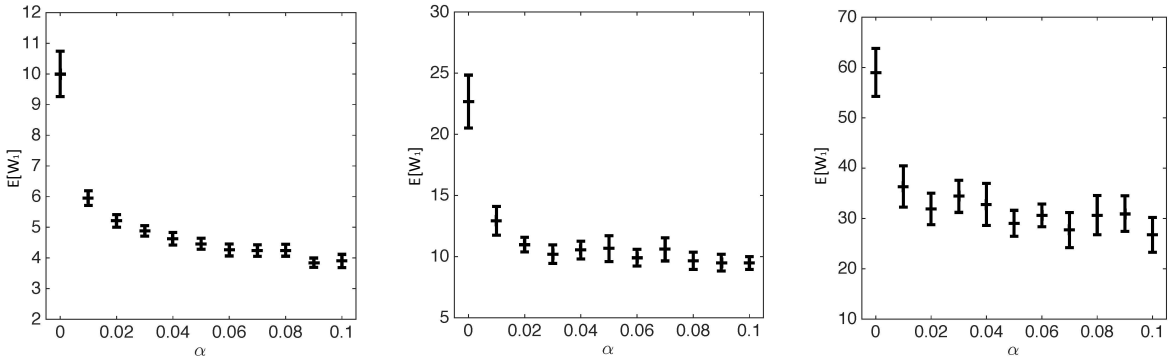


Figure 14 **Synchronization level as a function of α for different offered loads ($n_1 = n_2 = 25$, upper: $\rho = 0.95$, middle: $\rho = 0.9$, lower: $\rho = 0.85$).**

Synchronization leads to better load balancing, and therefore to better performance as measured by expected waiting times. Figure 15 plots the expected waiting time as a function of α for different values of load, ρ. We observe that the expected waiting time decreases with α, and most of the improvement is achieved with small values of α. In all load levels, the expected waiting time can be reduced by more than half if some patients act strategically.

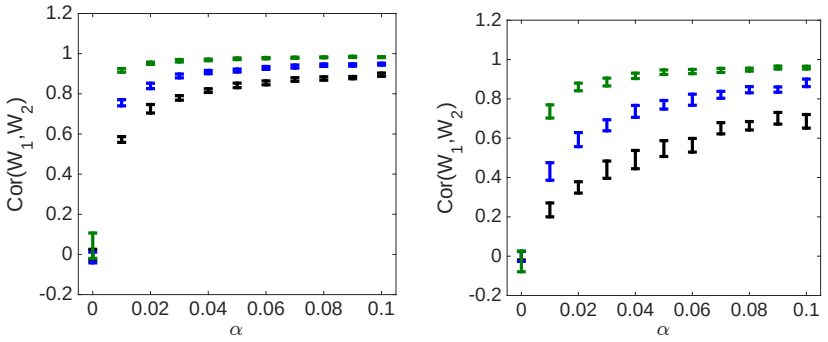


(a) $\rho = 0.85$ (b) $\rho = 0.9$ (c) $\rho = 0.95$

Figure 15 **Expected waiting time (in minutes) as a function of α for different offered loads ($n_1 = n_2 = 25$).**

Figure 16 plots synchronization levels as a function of α for different values of system scale, n_1, ($n_1 = n_2$). We observe that for the same values of α and ρ, the smaller system ($n_1 = n_2 = 10$) achieves greater synchronization compared to the larger system ($n_1 = n_2 = 40$). The intuition behind this

is that large systems are relatively well-balanced when acting alone, while small systems benefit more from the load balancing effect when connected to another system. More specifically, for an $M/M/n$ queue, the conditional waiting time, $W|W>0$, follows an exponential distribution with rate $n\mu(1-\rho)$ (Whitt 1992). Thus, for the same offered load ρ, the larger the system scale n, the smaller the mean and variance of the conditional waiting times. Therefore, when customers act strategically, we gain more benefit from the load balancing effect for small systems.



(a) $n_1=n_2=10$ (b) $n_1=n_2=40$

Figure 16 **Synchronization level as a function of α for different hospital sizes (upper: $\rho=0.95$, middle: $\rho=0.9$, lower: $\rho=0.85$).**

5.1.2. Non-symmetric case: The impact of patient preferences and ED size heterogeneity We next consider two non-symmetric cases—hospitals that differ in terms of patients' preferences (β_i) and hospitals that differ in the size (capacity) of their ED, expressed by a difference in n_i for $i=1,2$. The delay announcements themselves are still exact. We start by assuming that the EDs are the same size but that patients have a predetermined preference for the second hospital (i.e., $\beta_2>\beta_1$). (This can happen, for example, due to differences in the quality-of-care provided or constraints imposed by insurance providers that drive a higher proportion of the nearby population to a specific facility.) As a result, when $\alpha=0$, $p_1(w_1,w_2)=\exp(-\beta_1)/(\exp(-\beta_1)+\exp(-\beta_2))<0.5$. Hence, Hospital 2 has a higher demand than Hospital 1 to start with. Expressed mathematically, if $\rho_1=\lambda p_1(w_1,w_2)/(n_1\mu)$ and $\rho_2=\lambda p_2(w_1,w_2)/(n_2\mu)$, then $\rho_2>\rho_1$. We also notice that since

$$p_1(w_1,w_2)=\frac{1}{1+\exp((\beta_2-\beta_1)-\alpha(w_2-w_1))},$$

we can measure the heterogeneity in preferences by $|\beta_2-\beta_1|$.

Our numerical experiments lead to the following observations:

Observation 4. Preference heterogeneity reduces synchronization. Synchronization always leads to better performance of the more preferred (more loaded) queue, while performance of the

less preferred (less loaded) queue may rise or fall depending on the level of heterogeneity between them.

Figures 17 and 18 provide a numerical example in this setting. We define hospital Network A to have $\beta_1 = 1$ and $\beta_2 = 1.1$ and Network B to have $\beta_1 = 1$ and $\beta_2 = 2$. Both networks have the same total load ($\rho = \lambda/(\mu(n_1 + n_2)) = 0.9$), but the partition of the load between the hospitals in the network differs. In Network A,

$$\rho_1 = \lambda \frac{\exp(-\beta_1)}{\exp(-\beta_1) + \exp(-\beta_2)} \frac{1}{n_1 \mu} \approx 0.855$$

and

$$\rho_2 = \lambda \frac{\exp(-\beta_2)}{\exp(-\beta_1) + \exp(-\beta_2)} \frac{1}{n_2 \mu} \approx 0.945.$$

In system B, when $\alpha = 0$, $\rho_2 > 1$ (i.e., Hospital 2 is unstable when acting alone, so we start the plot of expected waiting times from $\alpha = 0.01$). In Figure 17, we observe that, as in the symmetric case, synchronization increases as α increases, and most of the load balancing effect is achieved with small values of α. For the same value of α, synchronization decreases with the level of asymmetry, $|\beta_2 - \beta_1|$ – networks that are more balanced in demand can achieve higher synchronization. Figure 18 shows that in both networks, the expected waiting time of the more loaded hospital decreases with α. Moreover, in Network B, synchronization assures stability. Nevertheless, this comes at the cost that w_1 increases slightly with α. In general, for small values of $|\beta_2 - \beta_1|$, the expected waiting time of the less loaded system falls with α, while for large values of $|\beta_2 - \beta_1|$, it rises slightly with α.

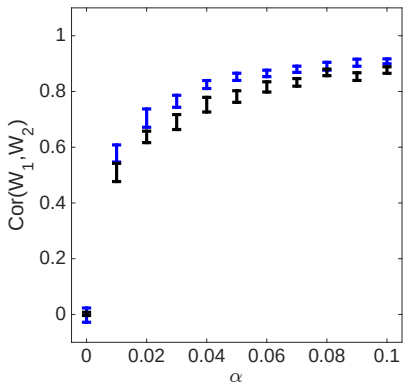
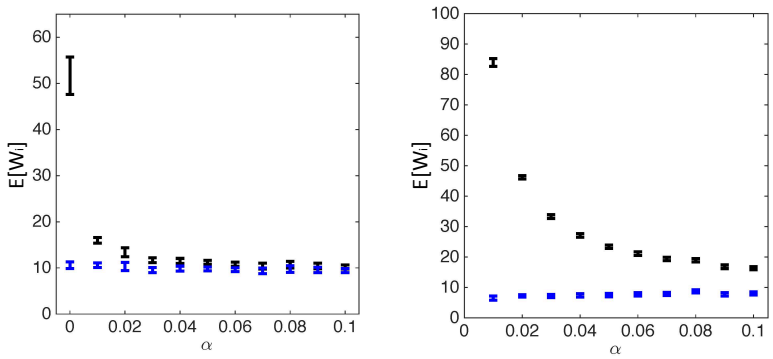


Figure 17 **Synchronization level as a function of α for different values of β_2 ($n_1 = n_2 = 25$, upper: Network A ($\beta_1 = 1$, $\beta_2 = 1.1$), lower: Network B ($\beta_1 = 1$, $\beta_2 = 2$)).**

We make similar observations when the heterogeneity is in staffing levels (n_1, n_2) (while all other parameters are equal). Synchronization increases with α, and the expected waiting time of the



(a) Network A ($\beta_1 = 1$, $\beta_2 = 1.1$) (b) Network B ($\beta_1 = 1$, $\beta_2 = 2$)

Figure 18 **Expected waiting time as a function of α for different values of β_2 ($n_1 = n_2 = 25$, upper: w_2, lower: w_1).**

more loaded hospital (the hospital with fewer beds/staff) falls with α. For a fixed value of α, synchronization decreases with $|n_2 - n_1|$. For small values of the level of asymmetry, $|n_2 - n_1|$, the expected waiting time of the less loaded system (the system with more beds/staff) decreases with α, while for large values of $|n_2 - n_1|$, the expected waiting time of the less loaded system increases with α. Here, again, synchronization has the potential benefit of ensuring stability.

5.2. The importance of timely and accurate delay announcements

The waiting time estimator used by hospitals will err from time to time. This kind of inaccuracy in delay estimators is inevitable. In this section, we show that these errors limit the degree of synchronization that can be achieved. We also show that delays of the wait time estimator in reflecting the true wait time may make the system more volatile (oscillating), and cause synchronization to fall (instead of rise) with α.

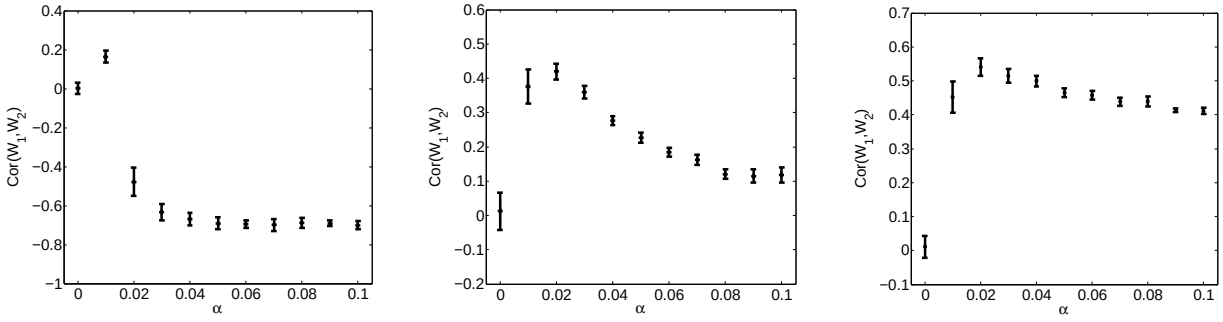
In practice, hospitals in the US publish different wait time estimators: All hospitals in our empirical study calculate waiting times by a moving average with a 4-hour window. Other websites and online apps report historical averages using even longer periods of time—up to 1 year (see, for example, the online app "ED Wait Watcher" (Groeger et al. 2014)). In recent years, a few hospitals have started employing waiting time estimators that are based on shorter time windows. For example, Stanford Hospital reported a moving average with a one-hour window. Our results show the potential effect of different wait time estimators, thus provide guidance on what delay announcements to use.

We compare the following two delay estimators:

1. *Moving average:* Historical average over time windows of specific lengths (the method currently used by most hospitals). Let l be the time window for the moving average function.

2. *Head-of-Line (HOL) wait*: Time waited by the customer who is currently at the head of the line when a new customer arrives. This method could be considered as a moving average with $l = 0$. This method was proved to be quite accurate for estimating delays in multi-server queuing systems (Ibrahim and Whitt 2009, Senderovich et al. 2014).

We start by simulating a network with two time-homogeneous queues with Poisson arrivals and Exponential service times as in Section 5.1. Figure 19 shows how synchronization levels change with the sensitivity parameter α for different window lengths l. Here, in contrast to what we observed with the idealistic model, synchronization first increases, but then decreases with α. When $l = 4$ hours, synchronization actually falls to *negative* levels, while when $l = 30$ minutes and HOL ($l = 0$), synchronization is positive for all values of α. Figure 20 shows how the expected waiting time changes with the sensitivity parameter α for different window lengths l. We observe that when the averaging window is long ($l = 4$ hours), the expected waiting time *increases* with α for moderate to large values of α. This suggests that the system is better off without announcements if patients are highly sensitive to delays. However, if the window used is small enough, the expected waiting time falls as sensitivity rises.

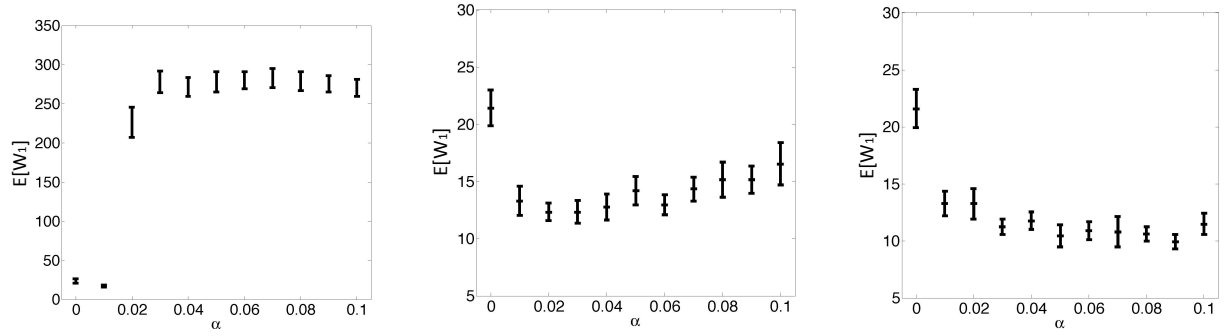


(a) $l = 4$ hours (b) $l = 30$ minutes (c) HOL Wait ($l = 0$)

Figure 19 Synchronization level as a function of α ($n_1 = n_2 = 25$, $\rho = 0.9$)

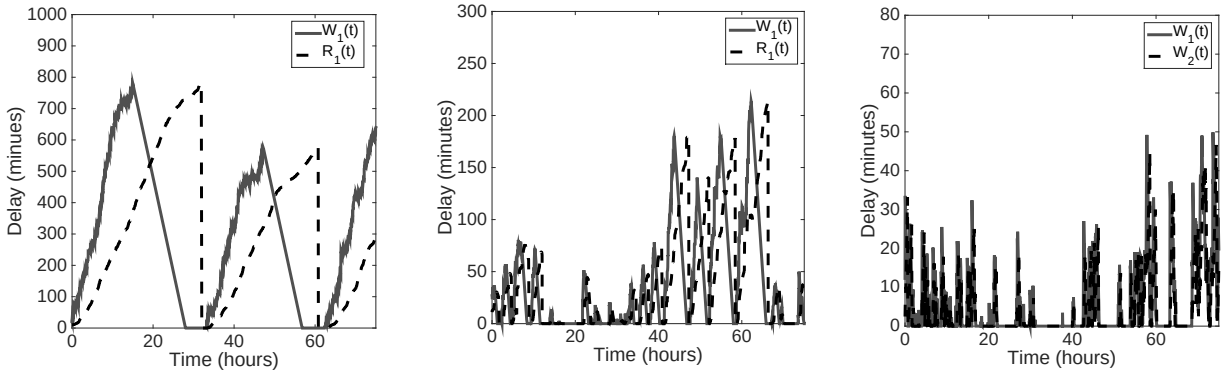
Our observations reflect differences in accuracy between the estimators. Indeed, the error rate of the reported waiting times with respect to the true waiting times increases with l. Specifically, when $\alpha = 0.1$, $l = 4$ hours, the root mean square error of the reported waiting times for Hospital 1, $\sqrt{E[(R_1 - W_1)^2]}$, is approximately 341.99; when $l = 30$ minutes, $\sqrt{E[(R_1 - W_1)^2]} \approx 32.81$, and when $l = 0$ (HOL), $\sqrt{E[(R_1 - W_1)^2]} \approx 12.17$. Figure 21 shows the sample path of the true waiting times versus the reported waiting times for different values of l, demonstrating the increase in inaccuracy is accompanied with a considerable time-lag between the two.

Interestingly, it is not the error alone which drives the phenomenon of performance deterioration with α. To validate this, we analyze a *modified-idealistic* model where we report the true waiting



(a) $l = 4$ hours (b) $l = 30$ minutes (c) HOL Wait ($l = 0$)

Figure 20 Expected waiting time as a function of α ($n_1 = n_2 = 25$, $\rho = 0.9$).



(a) $l = 4$ hours (b) $l = 30$ minutes (c) HOL ($l = 0$)

Figure 21 Sample path of the true and reported waiting times in hospital 1 ($\alpha = 0.1$, $n_1 = n_2 = 25$, $\rho = 0.9$).

time plus an error term that is normally distributed with mean 0 and standard deviation σ. Large values of σ lead to inaccurate delay announcements, but there is no delay effect as with the moving average method. Figure 22 shows how σ affects synchronization levels. We observe that synchronization increases monotonically with α for each value of σ, which is in contrast to the non-monotonic effect of α we observed in Figure 19. This is because although increasing α increases the error, the overestimates/underestimates are random, i.e. the reported waiting times are not distorted in a systematic way. Still, the inaccuracy has an impact—synchronization will not reach its full potential when the announcement is inaccurate. This is reasonable, as patients relying on inaccurate information may inadvertently choose the more loaded queue. Hence, as the error size rises, the maximal synchronization level will fall.

To understand why a moving average estimator performs so poorly, we next take a closer look at the sample path of the two queues for $\alpha = 0.1$ (see Figure 23). We observe that when $l = 4$ hours, 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

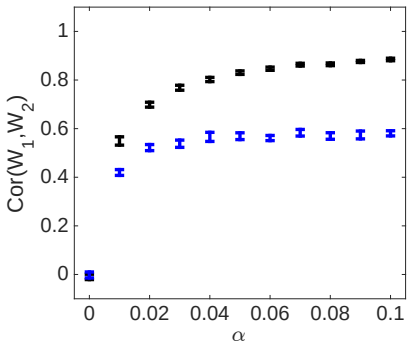
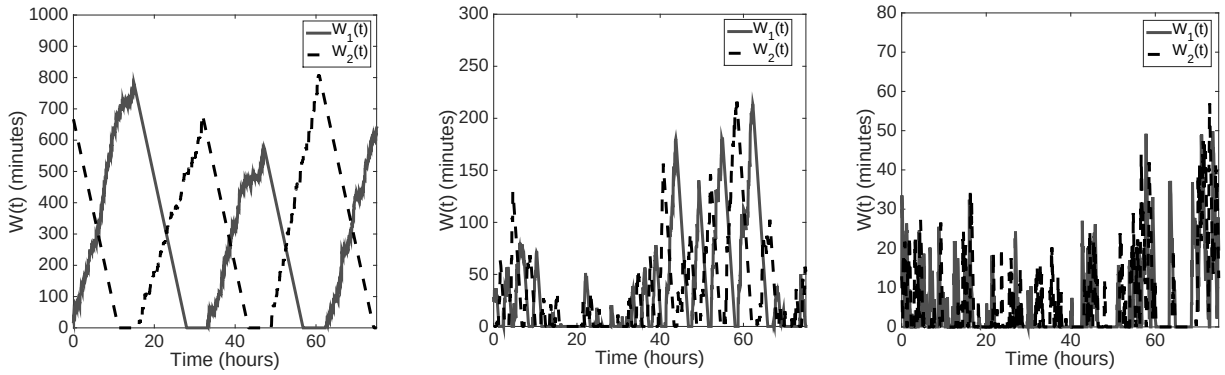


Figure 22 **Synchronization level for different values of α and σ (upper: $\sigma = 10$, lower: $\sigma = 100$, $n_1 = n_2 = 25$, $\rho = 0.9$).**

between the two waiting times. This phenomenon is known in control theory as *self-oscillation*, where systems with delayed feedback may oscillate solely because of the delay (Jenkins 2013). Here the delay announcement and its influence on customer choice can be considered a control mechanism. When $l = 30$ minutes or 0 (HOL), the two waiting time processes are closer to each other. This suggests that the delay effect is the main reason for the "desynchronization" when patients are very sensitive to delay. Figure 24 demonstrates that such oscillations are indeed observed in our data.



(a) $l = 4$ hours (b) $l = 30$ minutes (c) HOL ($l = 0$)

Figure 23 **Sample path of the waiting time process in the two hospitals ($\alpha = 0.1$, $n_1 = n_2 = 25$, $\rho = 0.9$).**

We next verify these results using more realistic settings as in Section 4. In particular, we assume time-varying Poisson arrivals with arrival rate function plotted in Figure 10(a), and time-varying staffing with two levels for day/night hours. We also assume i.i.d. service times following a log-normal distribution with a mean of 108 minutes.

Table 5 summarizes the simulation results for different values of l from 0 minutes (HOL) to 100 days. We observe that as l increases, the synchronization level decreases and the system performance

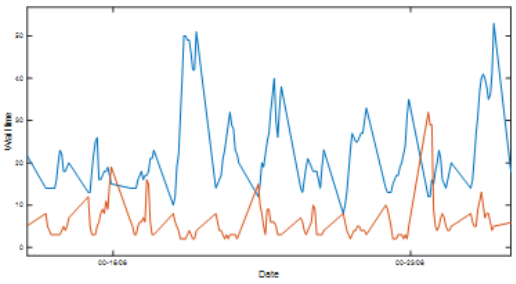


Figure 24 **Sample path of the waiting time data in the two oscillating hospitals.**

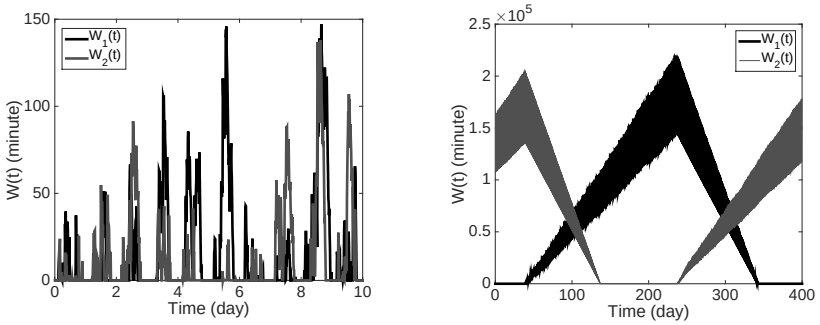
Table 5 **Synchronization level between the two systems and expected waiting times for different values of the time lag ($\theta = 1$, $\alpha = 0.05$)**

l	Cor	Cor (detrended)	$E[W_1]$	$E[W_2]$
0 (HOL)	0.45 ± 0.02	0.34 ± 0.02	18.55 ± 1.42	18.42 ± 1.33
30 minutes	0.17 ± 0.02	0.02 ± 0.02	22.46 ± 1.15	22.23 ± 1.04
4 hours	-0.09 ± 0.03	-0.21 ± 0.03	50.92 ± 5.11	51.16 ± 5.37
1 day	-0.71 ± 0.01	-0.81 ± 0.01	753.63 ± 22.73	746.84 ± 20.80
10 days	-0.81 ± 0.01	-0.90 ± 0.01	7447.42 ± 272.18	7670.28 ± 258.21
100 days	0.11 ± 0.05	-0.01 ± 0.04	26.75 ± 3.27	27.01 ± 3.78

deteriorates. The lowest wait times appear when using the HOL estimator. This is because the longer the time window, the more delay in feedback the system is suffering from. We observe an exception for the lag of 100 days. In fact, the intuition is that as the time lag grows beyond a certain level, the system is essentially reporting its long-run average performance. As the two systems are symmetric, they have the same long-run average performance. Thus, the delay information doesn't affect the system performance in this case. However, this is not entirely true. We did another simulation experiment for the lag of 100 days, but instead of starting both systems from empty, we start one system empty and the other system with a severely overloaded state. Then we see again the alternatively overloaded and underloaded trajectory of the waiting times of the two hospitals (Figure 25(b)). We conjecture that when l grows beyond a certain level, the system has two stable regions: In one region, the reported waiting times are the long-run average waiting times and the two hospitals are operating independently; in the other region, the two hospitals are alternating between being overloaded and underloaded. In reality, when l is very long, we would expect patients to disregard the real-time information value of the reported waiting times. Thus, the system is more likely to be in the first stable region.

We summarize the observations of this section as follows:

OBSERVATION 5. a) More accurate delay estimators lead to greater synchronization and better performance.



(a) Both empty (plot of 10 days)

(b) Hospital 1 empty, Hospital 2 with 100 patients waiting (plot of 400 days)

Figure 25 Sample path of the waiting time process in the two hospitals with different initial conditions ($\theta = 1$, $\alpha = 0.05$, $l = 100$ days)

b) A time-lag in the reported delay has a negative effect on synchronization and thus performance.

c) Large lags may desynchronize the system and lead to worse performance when patients are very sensitive to delay.

6. Discussion and concluding remarks

In this paper, we investigated the impact of ED delay announcements on patients' choice, and the effect of patients' choice on hospital synchronization and expected waiting times. We show that, even though hospitals consider delay announcements primarily a marketing tool, they also have an operational impact. Providing such information can improve patient safety at a very low cost. We provide empirical evidence that patients indeed take delay information into account when choosing an ED service, and that by providing such information, hospitals can significantly reduce waiting times. On the other hand, our data suggests that a risk exists. Some hospital pairs are negatively correlated, which suggests that the load in such an area alternates between hospitals. The regression tree suggests that this happens where the proportion of hospitals providing information is low, but the population seeking delay information through the internet is high.

Here the importance of the simulation model comes to the fore, explaining the mechanism by which such phenomena are created through the concept of delayed information. We calibrated the simulation model with realistic data and showed that the patient choice mechanism we suggested can reproduce phenomena observed in empirical data. Specifically, we showed that the range of network synchronization, including its negative part, appears in the simulation. We also found that a mismatch between delay announcements and actual delays may cause additional oscillations in the system load when patients are very sensitive to delays. Using numerical simulations on a simplified

version of the model, we observed that synchronization between systems increases with patients' sensitivity to waiting, the load of the system, and the accuracy of delay announcements. As a result, smaller and/or more loaded hospitals will benefit the most from such a policy. Furthermore, hospitals need not disseminate the availability of such information widely, as most of the advantages are achieved by small exposure rates.

With respect to the link between our empirical study and numerical experiments, we note that although the hospitals in our empirical study all use a moving average with a 4-hour window for delay announcements, the correlation of RWT between hospitals is still positive in general. This suggests that the number of patients who use this information when choosing which hospital to go to is currently relatively small. In this case, even with only a few patients acting strategically, we still gain improvements in performance as measured by expected waiting times. However, as the number of people who use this delay information grows, as evident from Figure 1(b), we will need better delay estimators to achieve optimal performance.

We note that hospitals have a good reason to choose a 4-hour moving average, as this provides a more stable estimator than the HOL. Nevertheless, the use of this estimator comes at a cost that needs to be considered. Specifically, our case study suggests that it limits wait time reduction to 16%. Hence, we recommend not using the 4-hour moving average if the proportion of strategic customers is large or the cost of waiting is large.

This paper opens several directions for future research. First, the instability caused by the mismatch between historical averages and future delays calls for more accurate machinery for delay announcements. We hypothesize that customers will be better served by estimates of future waiting times, given that wait time reports are accessed before patients travel to the ED and enter the system. Second, the delay until a physician is first seen represents only very partial information about the actual load in the ED. One might want to consider the influence of other information indicators such as ED LOS on patients' choices. Estimating future LOS is hard, as before a patient's arrival, the reason of his visit is unknown, and the treatment required, 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 go to. Lastly, we suggest that further investigation is needed to develop data and tools for estimating how patients perceive the cost of waiting in EDs. Such an estimation can help hospitals adjust their announcement policy and balance accuracy and stability. We are in the process of collecting such data, and hope to continue research in this direction.

7. Appendix

7.1. In-cluster RWT correlations

Figure 26 shows a histogram of the in-cluster RWT correlation values.

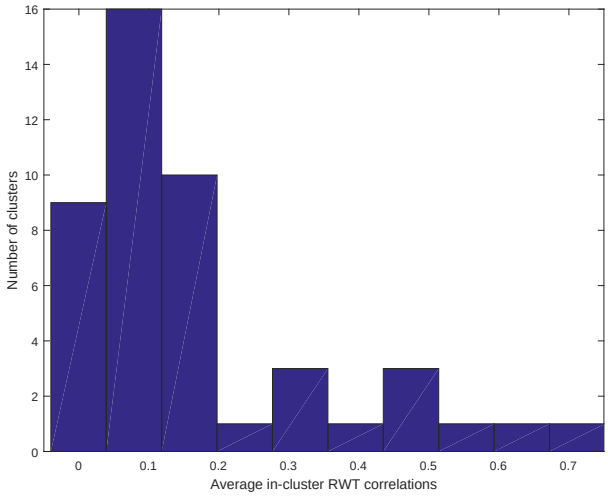


Figure 26 Histogram of in-cluster RWT correlation values.

Table 6 shows the correlation values between individual attributes and in-cluster RWT correlations.

Variable	Correlation	p-value
Number of reporting hospitals per square km	0.55	0.0001
Number of hospitals per square km	0.41	0.0044
Number of Bing queries	-0.236	N.S.
Number of Bing sessions with multiple hospitals	-0.118	N.S.
Number of Bing queries mentioning 'wait'	-0.09	N.S.
Number of pediatric hospitals	-0.11	N.S.
Queries per population	0.15	N.S.
Primary care people per 100k	-0.37	0.0114
Poverty level	0.05	N.S.
Household income	-0.30	0.04
Median male age	0.40	0.0056
Median female age	0.38	0.0098
Sex ratio	-0.29	0.0456
Fraction of children under 5 y.o.	-0.54	0.0001
Fraction of people over 65 y.o.	0.37	0.0079
Total number of people	-0.16	N.S.
Only wait data of the hospital shown	0.12	N.S.
Wait data of both hospital and adjacent hospitals shown	-0.28	0.0631
Wait data of adjacent hospitals shown after click	-0.10	N.S.

Table 6 Correlation values between individual attributes and in-cluster RWT correlations. "N.S." denotes not statistically significant at $P < 0.1$.

7.2. Robustness check - natural experiment

We model the effect of these breaks in wait time reporting using a linear regression model. Let A be the hospital for which a break in reporting occurred, and B be the closest hospital to A in our data set. The dependent variable in our model is the ratio between the average wait times in B in the 6 hours following the break in reporting in A, divided by the average wait times in B in the 6 hours prior to the break. If information is presented we expect the break to increase wait times in

hospital B, hence the ratio shall be larger than 1; if information is not presented, we expect the ratio to be close to 1, as practically there was no difference in the information given to patients in that case. Our independent variables are:

1. Average wait time at B in the 6 hours prior to the break, $\bar{T}_{-6:-1}$
2. Time of the day, h
3. Does hospital B display wait times? ($Disp_B$)
4. Does hospital A display wait times? ($Disp_A$)

The regression model is:

$$y = \beta_0 + \beta_1 \cdot \bar{T}_{-6:-1} + \beta_2 \cdot h + \beta_3 \cdot Disp_B + \beta_4 \cdot Disp_A + \epsilon$$

The results of this model are shown in Table 7. The statistically significant parameter is whether Hospital A displays wait times on its website. If Hospital A displays wait times ($Disp_A = 1$), withdrawing such information will affect wait times in the network. In particular, wait times after the breaks in Hospital B increase ($\beta_4 > 0$). This is because in comparison to before the break, Hospital A is not providing wait times, and therefore Hospitals A and B are less synchronized.

Variable	Parameter estimate (SE)	p-value
Wait time at B before break	0.03 (0.18)	0.87
Time of day	-0.16 (0.11)	0.16
Hospital B reports wait times	-5.60 (3.18)	0.08
Hospital A reports wait times	4.00 (1.84)	0.03

Table 7 Parameters of a regression model for the change in wait times at Hospital B, when the closest hospital A stops reporting wait times.

References

Abadie, A. 2005. Semiparametric difference-in-differences estimators. *Rev. Econom. Stud.* **72**(1) 1–19.

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, M., S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, G. Yom-Tov. 2015. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* **5**(1) 146–194.

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.

Bennidor, Raviv, Shlomo H. Israelit. 2015. Emergency department intermediate stay unit - a failed model. Technion working paper.

Campello, F., A. Ingolfsson, R.A. Shumsky. 2017. Queueing models of case manager. Management Science.

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

Centers for Disease Control and Prevention. 2010. National hospital ambulatory medical care survey (accessed 7/21/2015). URL `ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/dataset_documentation/nhamcs/stata/`.

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.

Ekström, Andreas, Lisa Kurland, Nasim Farrokhnia, Maaret Castrén, Martin Nordberg. 2014. Forecasting emergency department visits using internet data. *Annals of Emergency Medicine* **65**(4) 436–442. doi: 10.1016/j.annemergmed.2014.10.008.

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

Fox, Susannah, Maeve Duggan. 2013. Health online 2013. URL `http://www.pewinternet.org/2013/01/15/health-online-2013/`.

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

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

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.

Libby, Dan. 1999. Rss 0.91 spec, revision 3. URL https://web.archive.org/web/20001204093600/http://my.netscape.com/publish/formats/rss-spec-0.91.html.

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.

Meyer, Breed D. 1995. Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics* **13**(2) 151–161.

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

NEHI. 2010. A matter of urgency: Reducing emergency department overuse. New England Healthcare Institute, Research Brief.

Ofran, Yishai, Ora Paltiel, Dan Pelleg, Jacob M Rowe, Elad Yom-Tov. 2012. Patterns of information-seeking for cancer on the internet: an analysis of real world data. *PLoS One* **7**(9) e45921.

Perrin, Andrew, Maeve Duggan. 2015. Americans' internet access: 2000-2015. URL http://www.pewinternet.org/2015/06/26/americans-internet-access-2000-2015/.

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

Polgreen, Philip M, Yiling Chen, David M Pennock, Forrest D Nelson, Robert A Weinstein. 2008. Using internet searches for influenza surveillance. *Clinical Infectious Diseases* **47**(11) 1443–1448.

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

Selen, Jori, Ivo Adan, Stella Kapodistria, Johan S.H. van Leeuwaarden. 2016. Steady-state analysis of shortest expected delay routing.

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

Shi, P., M. C. Chou, J. G. Dai, D. Ding, J. Sim. 2015. Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science* **62**(1) 1–28.

Song, H., A.L. Tucker, K.L. Murrell. 2015. The diseconomies of queue pooling: an emprical investigation of emergency department length of stay. *Management Science* **61**(12) 3032–3053.

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.

Whitt, W. 1992. Understanding the efficiency of multi-server service systems. *Management Science* **38**(5) 708–723.

Widrow, Bernard, Istvan Kollar, Ming-Chang Liu. 1996. Statistical theory of quantization. *IEEE Transactions on Instrumentation and Measurement* **45**(2) 353–361.

Xu, Yuqian, Mor Armony, Anindya Ghose. 2016. The effect of online reviews on physician demand: A structural model of patient choice. NYU working paper.

Yom-Tov, Elad, danah M boyd. 2014. On the link between media coverage of anorexia and pro-anorexic practices on the web. *International Journal of Eating Disorders* **47**(2) 196–202.

Yom-Tov, Elad, Evgeniy Gabrilovich. 2013. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *Journal of Medical Internet Research* **15**(6).

Yom-Tov, Elad, Mounia Lalmas, Ricardo A. Baeza-Yates, Georges Dupret, Janette Lehmann, Pinar Donmez. 2013. Measuring inter-site engagement. *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*. 228–236.

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