

The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence

Gad Allon, Sarang Deo, Wuqin Lin

Kellogg School of Management, Northwestern University, Evanston, IL 60208

g-allon, s-deo, wuqin-lin@kellogg.northwestern.edu

In recent years, growth in the demand for emergency medical services along with decline in the number of hospitals with emergency departments (EDs) has raised concerns about the ability of the EDs to provide adequate service. Many EDs frequently report periods of overcrowding during which they are forced to divert incoming ambulances to neighboring hospitals, a phenomenon known as “ambulance diversion”. This paper aims to study the impact of key structural characteristics of the hospitals such as the number of ED beds, the number of inpatient beds, and the utilization of inpatient beds on the extent to which hospitals go on ambulance diversion. We first develop a simple queueing network model to describe the patient flow between the ED and the inpatient department and analyze this model using heavy traffic approximation. We show that, for a pre-specified delay probability, the fraction of time that the ED goes on diversion is decreasing in the spare capacity of the inpatient department and in the size of the ED, where both are appropriately normalized for the size of the inpatient department. We then test these findings by estimating a selection model using publicly available cross-sectional data on California hospitals and find moderate support for our theoretical findings. We also find evidence that certain hospitals, owing to their location and ownership structure, are more likely to choose ambulance diversion to mitigate overcrowding than others.

Key words: emergency department, empirical research, sample selection model, heavy traffic approximation

1. Introduction

Annual visits to the emergency departments (EDs) in the US increased by 18% from 1994 to 2004 due both to the growth in population and in per capita consumption of emergency medical services. In the same period, the number of hospitals operating 24 hour EDs declined by 12% (Burt and McCaig 2006). This has led to a growing number of hospitals reporting overcrowding situations in their EDs. One of the most important consequence of this overcrowding is the inability of the EDs to accept incoming ambulances for extended periods of time, a phenomenon known as ambulance diversion. A number of root causes, both within (shortage of ED beds and staff) as well as outside the ED (shortage of inpatient beds, delays in diagnostic services), have been proposed to explain ED

overcrowding and ambulance diversions. However, most of the discussion is limited to qualitative commentary and surveys (Derlet and Richards 2000, The Lewin Group 2002, GAO 2003, Burt and McCaig 2006) and lacks a rigorous analytical approach. The objective of this paper is to inform this discussion by developing a simplified theoretical model of patient flow between the ED and the inpatient department of a hospital and by empirically testing the key qualitative findings from the analysis of this model.

We develop a two-station queueing network model (See Hershey et al. (1981) and Cooper and Corcoran (1974) for early examples of application of queueing theory to hospitals) to analyze the flow of patients in the ED and the inpatient department of the hospital. Patients arrive to the ED either by ambulance or by self-transportation and are either discharged or admitted to one of the inpatient department after treatment. If inpatient beds are unavailable, these admitted patients continue to occupy beds in the ED thus blocking them for new arrivals. With the objective of meeting a pre-specified level of delay probability (probability that an arrival has to wait for a bed), the ED decides to go on ambulance diversion if the number of blocked beds increases beyond a threshold¹. We use this queueing model to study the impact of hospital capacity and the size of the ED on the fraction of time that the ED spends on diversion as a measure of ED overcrowding.

In order to facilitate analytical treatment, we decompose the model into two sub-models corresponding to the ED and the inpatient department with suitable reparameterization. We derive their performance measures in the heavy traffic regime, since these systems are expected to be large and extremely busy. We then use these qualitative insights to guide the formation of hypotheses for our empirical model. We formulate a selection model in which the EDs' decision to use ambulance diversion or not, and the hours on ambulance diversion conditional on this decision, are jointly estimated using publicly available data on California EDs and find moderate support for our hypotheses.

Our paper makes the following contributions to the literature on emergency department operations:

¹ Such policies are routinely followed in practice (Green 2002, Adams 2008) and are discussed in greater detail in Section 4.1.2.

1. Our use of heavy traffic analysis to derive empirically testable hypotheses is novel. It provides a theoretical basis for controlling the size of the hospital in a cross-sectional sample of hospitals with different levels of inpatient and ED size and inpatient utilization. We find that the fraction of time spent on ambulance diversion does not depend on the utilization of the inpatient department and the size of the ED per se, but on these measures suitably normalized for the size of the inpatient department.
2. To our knowledge, this is the first study to highlight theoretically and then document empirically the **interaction** between the utilization of the inpatient department and the size of the ED as a key factor in contributing to ambulance diversion, i.e., we find that the magnitude of the marginal impact of the utilization of inpatient beds on the extent of diversion is decreasing in the size of the ED and vice versa.
3. As a consequence of this interaction, we find that the fraction of time spent on ambulance diversion (counter to intuition and empirical findings from single hospital studies) does not necessarily increase in the utilization of the inpatient department. In particular, if the size of the ED relative to the inpatient department is sufficiently small, the fraction of time spent on diversion is insensitive to the utilization of the in-patient department . This result suggests that it is not possible to directly generalize the findings from the single hospital studies to a larger population.
4. Results of the empirical study also suggest that our queuing network model, although stylized, captures the essential elements of patient flow that are pertinent in determining the extent of ambulance diversion in hospitals.

More broadly, this paper contributes to the study of ED operations in operations management and emergency medicine literature from both theoretical and empirical perspectives. The queueing network model spanning the emergency department (ED) and the inpatient department contributes to the theoretical literature, which has traditionally focused either outside the hospital system (ambulance planning) or within the ED (staffing and bed planning). Theoretical analysis of this network comprising the ED and the inpatient department allows us to quantify the impact of

main drivers for ambulance diversions, which have been previously identified in the emergency medicine literature. Our empirical analysis also contributes to the scant literature on the empirical estimation of queueing systems in the operations management literature.

The remainder of this paper is organized as follows. We provide a brief background on ED overcrowding and ambulance diversion in Section 2. The relevant literature is reviewed in Section 3. The queueing model and the related analysis are described in Section 4. The empirical model and the related analysis are described in Section 5. Section 6 contains concluding remarks. Proofs of all the theoretical results are provided in Appendix B.

2. Background

The phenomenon of ED overcrowding has received increasing attention in the popular press in the past decade (Shute and Marcus 2001, Gosselin 2001). In a national survey of ED directors, 91% of the respondents reported overcrowding as a problem with 39% reporting it as a daily occurrence (Derlet et al. 2001). While ED overcrowding can be intuitively understood as the imbalance between demand and the available capacity, there is no consensus on its precise definition. A number of indicators have been used to assess its extent: the time patients wait to receive service (waiting time), the total time spent by patients in the ED (length of stay), the percentage of patients who leave without being seen, the number of patients who remain in the ED after the decision is made to admit or transfer them (boarding patients), and the number of hours for which incoming ambulances are diverted to other hospitals (ambulance diversion) (Derlet et al. 2001, Burt and McCaig 2006, GAO 2003). In this paper, we focus on ambulance diversion as the key indicator of ED overcrowding because of its serious impact on various aspects of the public health system and its widespread prevalence.

Ambulance diversion is a phenomenon wherein an ED, reckoning that it does not have the requisite capacity to promptly care for additional emergency patients, informs the emergency medical service (EMS) providers to reroute incoming ambulances to other hospitals in its vicinity². Ambu-

²Several federal and local statutes aim to regulate the use of ambulance diversion by hospitals such as under what conditions and for how long can an ED go on diversionary status and who needs to be informed of the diversionary status.

lance diversion results in increased transit time for patients (Schull et al. 2003b), which increases the risk of poorer patient outcomes for certain conditions (Schull et al. 2004). It also reduces the responsiveness of EMS agencies as ambulances spend more time to find an open ED and/or to wait till the ED staff can accept the patients (Eckstein and Chan 2004, Kennedy et al. 2004). Moreover, ambulance diversion also results in substantial loss of revenue for hospitals since around 40% of all hospital admissions come through the ED (Melnick et al. 2004, Merrill and Elixhauser 2005). Despite these serious consequences, ambulance diversion is highly prevalent in the US: nearly half of all the hospitals reported time on diversion in 2004 (American Hospital Association 2005).

Ambulance diversion is also distinct from other indicators in that it is a decision made by the ED management while most other indicators are consequences of this decision. Hospitals reporting 20% or greater time on diversion in a survey had longer wait times for treatment, longer average lengths of stay and longer wait times for transfer from ED to a psychiatric bed (The Lewin Group 2002).

3. Literature Review

There is a vast literature on various aspects of managing ambulance service operations including fleet sizing (Savas 1969), location planning (Swoveland et al. 1973) and dispatching / deployment (Fitzsimmons 1973). See Green and Kolesar (2004) and references therein for a more complete list. However, almost all of this literature takes the perspective of an EMS agency while we take the perspective of a hospital or an ED.

Considerable work has been done on capacity management in EDs and inpatient department with the objective of meeting a certain delay performance criterion. Green et al. (2006) divide the workday into independent staffing periods and then employ an M/M/s model to derive the staffing level for each period so as to meet the desired service target such as probability of delay. Vassilopoulos (1985a) uses a dynamic programming formulation to decide on an hourly allocation of doctors that is proportional to the patient arrival rate. Green and Nguyen (2001), Green (2002) use simple M/M/s model to investigate the impact of inpatient occupancy rate on the probability

of delay in obtaining a bed and show that smaller hospitals need to maintain lower occupancy than larger hospitals in order to guarantee the same delay probability. However, these papers do not explicitly model the interaction between the ED and the inpatient department in the form of patient flow and the impact of inpatient capacity on ambulance diversion and delay performance in the ED.

Vassilocopoulos (1985b) studies the problem of allocating beds among inpatient departments so as to simultaneously satisfy several performance measures, including immediate admission of emergency patients. Gerchak et al. (1996) consider the allocation of operating room capacity between elective and emergency procedures with the objective of minimizing total cost of operation including overtime when emergency patients cannot be turned away. While these papers model the interaction between the ED and the inpatient department, they do not consider the case of ambulance diversion, i.e., emergency patients being turned away, which is the focus of our work. Moreover, the analytical approach used in these papers is substantially different from ours and they do not empirically validate their results.

There is limited literature on empirical estimation of queueing models. Joskow (1980) models the hospital as an $M/M/s$ queue where s corresponds to the number of beds. Using normal approximation for the delay probability, he shows that the average reserve margin of the hospital (number of beds less the average occupancy) varies proportional to the square-root of the average hospital occupancy. He then estimates this reserve margin as a function of the average occupancy and various measures of market competition and regulation. Mulligan (1985) relaxes several assumptions in the theoretical model of Joskow (1980) including the infinite buffer assumption. Ramdas and Williams (2008) examine the tradeoff between aircraft utilization and on-time performance for US airlines using a queueing paradigm. In conformance with the predictions from queueing models, they find that utilization of aircraft is negatively associated with on-time performance and that this effect is more negative for aircrafts with higher variability in their routes.

Our empirical approach is different from these papers in many significant ways. First, the formulation of the empirical model is guided by the theoretical analysis of the underlying queueing

network in the heavy traffic regime. This enables us to study the impact of the scale of the system on ambulance diversion while tackling the analytical complexities of our model. Second, we allow for the possibility that hospital managers may not use queueing rationale in deciding whether to divert incoming ambulances or not. We use a selection model to endogenize EDs' decision to go on diversion and use the queueing framework to estimate the extent of diversion conditional on this decision.

Given their significance, ED overcrowding and ambulance diversion are among the most actively researched issues in emergency medicine. Asplin et al. (2003) present a conceptual framework that partitions the ED system into three interdependent components: input (patient arrival), throughput (ED operations) and output (patient disposal including inpatient admission) and Soldberg et al. (2003) report an accompanying comprehensive list of performance measures across the three components. However, the list is inadequate to inform research questions since it precludes any hypothesis of how multiple measures might be correlated. For instance, extent of ambulance diversion, average number of ED beds blocked by boarded patients and average hospital occupancy were all rated among the most important measures in the output component. Our queueing model attempts to present a stylized formal representation of this framework and provide the requisite theory to link important measures in each component and derive empirically testable hypotheses.

Recent empirical studies in this literature have focused on either the analysis of time-series data from single hospitals or surveys of multiple hospitals. Using the data from one ED in Toronto, Schull et al. (2003a) found that the number of boarded patients in the ED was associated with duration of ambulance diversion episodes. Han et al. (2007) found that adding ED beds did not reduce the ambulance diversion hours in an urban, academic trauma center. In contrast, McConnell et al. (2005) found that increasing the number of ICU beds at an academic acute care facility decreased the time spent on ambulance diversion and reduced the ED length of stay for ICU patients. Similarly Forster et al. (2003) found that higher hospital occupancy was associated with shorter length of stay in the ED at an acute care teaching hospital. While these single location studies contribute to our understanding of the causes of ambulance diversion, the generalizability

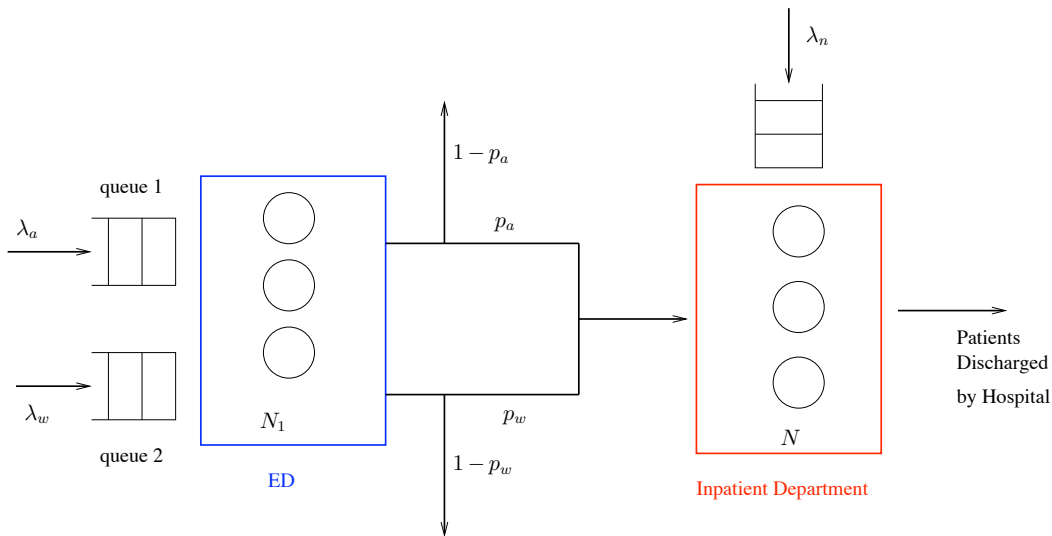
of their results to a larger population of EDs is unclear. Moreover, by design, these studies are not capable of testing the impact of structural characteristics such as the size, type and location of the hospital with ambulance diversion. We partly overcome these limitations by conducting our empirical analysis using a cross-sectional sample.

Burt and McCaig (2006) also use a cross-sectional sample and find that the time spent by hospitals on diversion increases with size. This runs counter to the well-known principle of statistical economies of scale in systems susceptible to congestion. According to this principle, larger systems tend to provide better delay performance to their customers than smaller systems for similar levels of utilization. This implies that larger hospitals, everything else being the same, should spend less time on ambulance diversion. We reconcile these two seemingly disparate observations by employing heavy traffic approximation of the queueing model, which allows us to appropriately redefine the measures of size and occupancy by explicitly accounting for the size of the hospital.

4. Model

Our primary objective in this paper is to gain an understanding of the relationship between the key structural components of the hospital and the extent of ambulance diversion that is empirically testable. As a result, given the aggregate nature of our data (described later) and foreseeable analytical complexity, we formulate a simplified steady-state stationary queueing network model to capture the essential elements of the patient flow in the ED and the inpatient department; we do not claim that the model can be used to accurately simulate the detailed dynamics of the ED patient flow. We show later, in our empirical study, that this simplified model does capture significant determinants of the extent of ambulance diversion at an aggregate level.

For our simplified model, we focus on the delay probability (the probability that an arrival to the ED has to wait for a bed) and the fraction of time that the ED goes on ambulance diversion as the performance measures of interest. In the short term, when the staff and bed capacity in the ED and the inpatient department are fixed, there is a potential tradeoff between these two performance measures. A reduction in the time on diversion might result in more arrivals, thereby increasing

Figure 1 A Two-Station Queueing Model

the congestion for arriving patients as reflected in a higher delay probability. Thus, we implicitly formulate the hospital's problem as choosing the extent of diversion so as to meet a pre-specified delay probability, which is a proxy for the ED's mission of providing timely care.

We first formulate a two-station queueing network (corresponding to the ED and the inpatient department), which is not amenable to exact analysis. We then simplify this network model by applying a series of approximations and then analyze the simplified model using heavy traffic approximations. We also conduct a simulation study to evaluate the quality of our approximations both within and without the heavy traffic regime.

4.1. A two-station Queueing Model

Figure 1 depicts the model of patient flow in the ED and the inpatient department of the hospital.

4.1.1. Emergency Department. We model the ED as a multi-server station where N_1 denotes the number of beds in the ED. Patients arrive to the ED either by ambulance (ambulance patients) or by self-transportation (walk-in patients) according to a Poisson process at the rate of λ_a and λ_w , respectively, and join two separate queues as shown in Figure 1. We assume that the ambulance patients receive non-preemptive priority over the walk-in patients: An empty ED bed is never allocated to a walk-in patient if there is an ambulance patient waiting. However, a patient cannot be removed from her bed during treatment. We also assume that the service time of each

patient (irrespective of the mode of arrival) in the ED is exponentially distributed with mean m_1 (see Green and Nguyen (2001) for a discussion of this assumption). After being treated in the ED, a fraction p_a of the ambulance patients and p_w of the walk-in patients are admitted to the inpatient department for further treatment and the remaining patients are discharged from the hospital.

4.1.2. Inpatient Department. We model the inpatient department as a multi-server station where N denotes the number of inpatient beds. The inpatient department receives two streams of arrivals: emergency patients who are admitted from the ED and non-emergency patients who arrive directly according to a Poisson process at the rate of λ_n where the former receive priority over the latter. If all inpatient beds are occupied, we assume that the non-emergency patients join a queue whereas the emergency patients continues to occupy ED beds (called boarding patients) since there is typically no waiting room between the ED and the inpatient department. Anecdotal evidence suggests that hospitals use a threshold on the number of boarding patients to formulate their ambulance diversion policy. For instance, Columbia Presbyterian Hospital in New York City goes on diversion when 15 or more patients are boarding (Green 2002) whereas Northwestern Memorial Hospital in Chicago goes on diversion when 14 or more patients are boarding (Adams 2008). Hence, we assume that the hospital goes on ambulance diversion if there are more than K boarding patients in the ED. Similar to the ED, we assume that the length of stay of each patient (emergency as well as non-emergency) in the inpatient department is exponentially distributed with mean m_2 .

4.2. A Simplified Queueing Model

The queueing dynamics of the two-station queueing network described above can be characterized by a Markov process that is quite complex. The steady state distribution of this process is difficult to derive and there is no simple closed form formula for either the delay probability of emergency patients or the fraction of time the ED is on diversion. Hence, we introduce the following approximations to the queueing network in order to improve analytical tractability:

- We approximate the number of available beds in the ED by $N_1 - B$, where B is the average number of blocked beds (or equivalently boarding patients) in the ED.

- We approximate the overall arrival process to the ED by a Poisson process having rate $\lambda_w + (1 - \delta)\lambda_a$, where δ is the fraction of time on diversion. While the arrival process to the ED is not Poisson due to ambulance diversion, such approximation can be justified when $\delta\lambda_a \ll \lambda_w$. This condition is easily satisfied for U.S. hospitals since around 15% of the arrivals are ambulance patients and the percentage of time on diversion is less than 10% (Burt and McCaig 2006, GAO 2003).

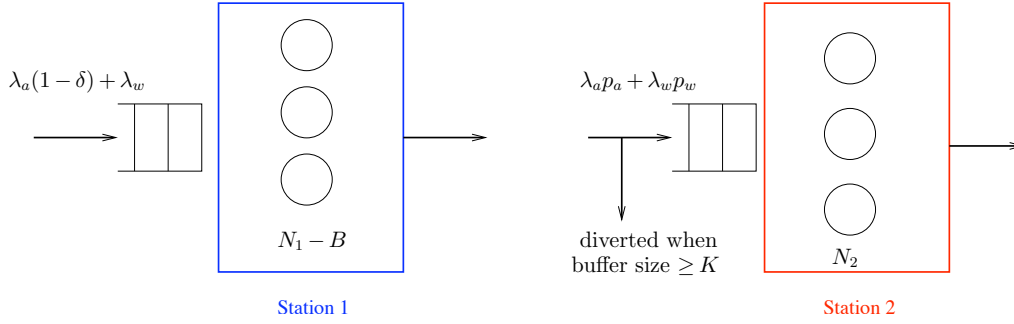
- We approximate the number of inpatient beds dedicated to emergency patients by $N_2 = \frac{\lambda_E}{\lambda_E + \lambda_n} N$ where $\lambda_E = \lambda_a(1 - \delta)p_a + \lambda_w p_w$ is the rate at which emergency patients are admitted to the inpatient department. Thus this approximation practically divides the inpatient bed capacity into two separate pools for emergency and non-emergency patients in the proportion of the arrival rate of the respective streams.

- We approximate the arrival of the emergency patients to the inpatient department by a Poisson process with rate $\lambda_a p_a + \lambda_w p_w$. This approximation will work well when $\lambda_a(1 - \delta)p_a \gg \lambda_w p_w$, i.e., when the majority of emergency patients admitted to the inpatient department are ambulance arrivals. This condition seems reasonable for U.S. hospitals since around 40% of the ambulance arrivals and around 10% of the walk-in patients are admitted on average (Burt et al. 2006).

Using these approximations the two-station queueing network in Figure 1 is reduced to two separate single-station queueing systems, as shown in Figure 2, where the ED is approximated as an M/M/($N_1 - B$) system (station 1) and the inpatient department is approximated by an M/M/ N_2/K system (station 2). In spite of these several simplifying approximations, we obtain very good estimates of the key performance measures, namely the fraction of time on diversion and the delay probability, as is evident from the results of the simulation study reported in Section 4.4 below.

For this simplified system, let $Q_1(t)$ and $Q_2(t)$ denote the number of patients (queue length process) at time $t > 0$ for station 1 and station 2 respectively. Then, the steady state delay probability and the long run fraction of time on diversion can be approximated, respectively, by

$$P_d = \mathbb{P}(Q_1(\infty) \geq N_1) \text{ and } \delta = \mathbb{P}(Q_2(\infty) \geq N_2 + K). \quad (1)$$

Figure 2 A Simplified Queueing Model

where $Q_1(\infty)$ and $Q_2(\infty)$ are the steady state queue lengths for station 1 and station 2 respectively. Similarly, the average number of blocked ED beds in steady state can be approximated by $B = \mathbb{E}[(Q_2(\infty) - N_2)^+]$.

4.3. Heavy Traffic Approximation

While we can estimate P_d and δ numerically using (1), it is not possible to provide their analytical characterization even for the simplified system. In order to achieve this, we next approximate $Q_1(\infty)$ and $Q_2(\infty)$ by their heavy traffic limits. As the name suggests, this approach involves approximating the original queueing system by a limit of a sequence of systems that approach heavy traffic (i.e., traffic intensity approaches 1). The benefit of this approach is that the performance measures of the limit system can be characterized analytically and they are good approximations for the performance measures of the system of interest provided its traffic intensity is sufficiently close to 1.

4.3.1. Basic approach. To illustrate the approach, consider an M/M/s system with arrival rate λ and service rate μ for each server. To derive the heavy traffic limit, we consider a sequence of M/M/n systems, each with service rate μ and indexed by $n = 1, 2, \dots$, that satisfies the following conditions:

(A1) For the n^{th} system, the number of servers is n and the arrival rate is λ^n , where the superscript n denotes the n^{th} system

(A2) $\lambda^n = \lambda$, i.e., the n^{th} system in the sequence is the original system, and

(A3) $\sqrt{n}(1 - \rho^n) \rightarrow \beta$ for some constant β as $n \rightarrow \infty$, where $\rho^n = \lambda^n / (n\mu)$ is the traffic intensity of the n^{th} system.

Condition (A3) is critical and needs more explanation. It states that for large systems, the excess capacity $(1 - \rho)$ should be approximately inversely proportional to the square-root of the number of servers. It is not only consistent with the common wisdom that the appropriate level of server utilization in service systems should increase with the size of the system but also further specifies that this increase at the rate proportional to the square root of the system size. This condition is also the theoretical underpinning for the famous “square-root” staffing rule that is used in call center management.

Halfin and Whitt (1981) provide theoretical justification for (A3) by showing that the probability of delay is approximately equal for each of the systems in the sequence if and only if this condition is satisfied. If the fraction of excess capacity goes down at a rate faster than $1/\sqrt{n}$, then an arriving customer has to wait almost surely. On the other hand, if the fraction of excess capacity goes down at a rate slower than $1/\sqrt{n}$, an arriving customer almost never waits.

Halfin and Whitt (1981) further show that under condition (A3), in steady state, the normalized queue length $\tilde{Q}^n(\infty) = (Q^n(\infty) - n)/\sqrt{n}$ converges to a diffusion process $\tilde{Q}(\infty)$ in distribution. Therefore, we can approximate $Q^n(\infty)$ by $\sqrt{n}\tilde{Q}(\infty) + n$, a property that will use extensively to characterize the performance measures for the system of our interest.

Similar approach can be adopted for an M/M/s/K system (Whitt 2004) by considering the limit of a sequence of systems, each with service rate μ and indexed by n such that:

(B1) The n^{th} system is an M/M/n/Kⁿ system with arrival rate λ^n

(B2) $\lambda^s = \lambda$ and $K^s = K$, i.e., the s^{th} system in the sequence is the original system, and

(B3) $\sqrt{n}(1 - \rho^n) \rightarrow \beta$ and $K^n/\sqrt{n} \rightarrow \kappa$ for some constants β and κ as $n \rightarrow \infty$.

4.3.2. Application to the hospital model. Using the above approach, we approximate the queue length processes Q_1 and Q_2 introduced in Section 4.2, by two diffusion processes \tilde{Q}_1 and \tilde{Q}_2 and use these processes to obtain an approximation for the delay probability P_d and the percentage of diversion hours δ for the system of our interest. Interested readers are referred to Halfin and Whitt (1981) and Whitt (2004) for the precise description of \tilde{Q}_1 and \tilde{Q}_2 .

We first consider station 2 (M/M/ N_2 / K system), which is the approximation of the inpatient department. Using condition (B3), we set $\beta_2 = (1 - \rho_2)\sqrt{N_2}$, $\kappa = K/\sqrt{N_2}$, where $\rho_2 = (\lambda_a p_a + \lambda_w p_w)m_2/N_2$ is the traffic intensity. Using (7.5) in Whitt (2004), we can approximate the fraction of time on diversion δ by

$$\tilde{\delta}(N_2, \beta_2, \kappa) = \frac{\beta_2 e^{-\kappa\beta_2}}{(\sqrt{N_2} - \beta_2) \left(1 - e^{-\kappa\beta_2} + \beta_2 \frac{\Phi(\beta_2)}{\phi(\beta_2)} \right)}, \quad (2)$$

Below, we summarize some important structural properties regarding the function $\tilde{\delta}(N_2, \beta_2, \kappa)$, which we will use for our empirical analysis:

PROPOSITION 1. *The function $\tilde{\delta}$ satisfies:*

- (i) *If $K\rho_2 \geq 2$, $\tilde{\delta}(N_2, \beta_2, \kappa)$ is decreasing in β_2 .*
- (ii) *$\tilde{\delta}(N_2, \beta_2, \kappa)$ is decreasing in κ .*

Note that the condition in result (i) is easily satisfied for $\rho_2 \geq 0.8$ and $K \geq 3$, which is reasonable for most hospitals. Thus, result (i) states that the fraction of time the ED is on diversion is decreasing in the excess capacity of the inpatient department, normalized for its size. Similarly, result (ii) implies that increasing the diversion threshold K (hence κ) will reduce the percentage of time on diversion δ . However, as discussed earlier, increasing κ might also result in more congestion in the ED and increase the likelihood that an incoming patient has to wait for a bed (delay probability, P_d). Thus hospitals face a trade-off between two critical performance measures, the fraction of time on diversion and the delay probability.

In our model, we assume that the ED chooses the appropriate level of diversion threshold K^* to minimize the fraction of time on diversion while maintaining a certain level of delay probability \bar{P}_d since its primary mission is to provide timely service to its arriving patients (Green 2002, Green and Nguyen 2001). Hence, in order to characterize K^* , we first need to characterize how the delay probability depends on the diversion threshold K (or equivalently κ).

For this, consider station 1 (M/M/ $N_1 - B$ system), which is the approximation of the ED. Using Proposition 1 of Halfin and Whitt (1981) we approximate the delay probability as

$$\tilde{P}_d = \left[1 + \frac{\tilde{\beta}_1 \Phi(\tilde{\beta}_1)}{\phi(\tilde{\beta}_1)} \right]^{-1}, \quad (3)$$

where $\tilde{\beta}_1$ depends on N_1, N_2, β_2 , and κ . Since the expression is quite complicated, it has been relegated to Appendix A for the ease of exposition. In what follows, we use the notation $\tilde{P}_d(N_1, N_2, \beta_2, \kappa)$ so as to emphasize \tilde{P}_d as a function of N_1, N_2, β_2 and κ . The next Proposition formalizes the intuition that higher κ yields higher delay probability P_d .

PROPOSITION 2. *The function $\tilde{P}_d(N_1, N_2, \beta_2, \kappa)$ is increasing in κ .*

Proposition 2 implies that for any N_1, N_2, β_2 there is a unique κ that solves $\tilde{P}_d(N_1, N_2, \beta_2, \kappa) = \bar{P}_d$. Denote the solution by $\kappa^*(N_1, N_2, \beta_2, \bar{P}_d)$. We can then approximate the appropriate level of diversion threshold by $\tilde{K}^*(N_1, N_2, \beta_2, \bar{P}_d) = \sqrt{N_2} \kappa^*(N_1, N_2, \beta_2, \bar{P}_d)$.

PROPOSITION 3. *The function $\tilde{K}^*(N_1, N_2, \beta_2, \bar{P}_d)$ is increasing in N_1 .*

Proposition 3 is quite intuitive and implies that, everything else being equal, a hospital with larger ED will set a higher diversion threshold in terms of the number of boarding patients. Using Proposition 3 and result(ii) of Proposition 1, it is easy to see that the fraction of time on diversion is decreasing in $\frac{N_1}{\sqrt{N_2}}$.

To summarize, the analysis of our queueing network model using heavy traffic approximation allows us to formulate the following hypotheses regarding how the fraction of time spent by the ED on diversion depends on the key structural characteristics of the system:

1. The fraction of time spent on diversion is decreasing in the spare capacity of the inpatient department, appropriately normalized for its size (β_2), as observed from result (i) of Proposition 1.
2. The fraction of time spent on diversion is decreasing in the size of the ED, appropriately normalized for the size of the inpatient department ($N_1/\sqrt{N_2}$), as observed from Proposition 3 and result(ii) of Proposition 1.
3. There is an interaction between the two explanatory variables mentioned above, as observed from (2).

Table 1 Estimation of fraction of time on diversion: approximation vs simulation

N	ρ_2	Approximation	Simulation	Difference
35	0.98	0.103	0.107	3.44%
40	0.86	0.052	0.055	4.61%
45	0.76	0.024	0.0258	5.59%
50	0.69	0.0104	0.0115	8.86%
55	0.62	0.00423	0.0049	13.5%
60	0.57	0.00163	0.0019	14.1%

In Section 5, we test these hypotheses using a dataset of hospitals licensed in the state of California. However, before turning to the empirical study, we test the accuracy of our approximations in predicting the fraction of time on diversion and the delay probability using discrete event simulation.

4.4. Simulation

While the primary purpose of the simulation is to validate the accuracy of the estimation of fraction of time on diversion, we also validate the accuracy of the delay probability estimation. We compute the fraction of time on diversion using the heavy traffic approximation and a discrete-event simulation tool where we vary the size of the hospital and fix the other parameters. In the validated setting, we assume that $K = 4$, and $N_1 = 10$, the arrival rates are $\lambda_a = 6$, $\lambda_w = 15$, and $\lambda_n = 5$, the service times are $m_1 = 0.5$, and $m_2 = 3.5$. We also assume that the likelihood that customers arriving to the emergency department continue to an inpatient department are $p_a = 0.8$ and $p_w = 0.005$. Table 1 shows the estimates of fraction of time on diversion using both heavy traffic approximation and simulation, and the difference between along with the estimated traffic intensity level of the inpatient department.

First, observe that the accuracy of the approximation is extremely high as long as the traffic intensity of the hospital is high enough. Second, as we fix the arrival rates and the size of the ED, the fraction of time on diversion decreases with increasing number of inpatient beds, as estimated by the simulation and computed by the approximation.

Table 2 Estimation delay probability: approximation vs simulation

N_1	ρ_1	Approximation	Simulation	Difference
9	0.97	0.89	0.91	2.1%
10	0.86	0.57	0.59	3.5%
11	0.77	0.36	0.37	4.7%
12	0.70	0.23	0.25	5.4%
13	0.65	0.15	0.16	6.1 %
14	0.60	0.09	0.10	7.3 %
15	0.55	0.05	0.07	11.9%

Next, we validate the accuracy of the delay probability estimates via the heavy traffic approximation relative to a simulation-based estimation. We fix the parameter as described above. In addition, we fix the number of inpatient beds to $N = 35$. Table 2 reports, for different sizes of the ED, the delay probability as computed by the heavy traffic approximation and a discrete-event simulation, as well as the difference between the two estimates. The table also reports the traffic intensity of the emergency department. Again, observe that the accuracy of the delay probability estimate increases as the traffic intensity of the emergency department increases. These results suggest that the series of approximations introduced in Section 4 and the heavy traffic approximation provide a very good description of the actual dynamics of the original network.

5. Empirical Analysis

Our analysis of the queueing network model using heavy traffic approximation in section 4 suggests that the fraction of time spent on diversion is decreasing in the size of the ED and the spare capacity in the inpatient department, suitably adjusted for the size of the inpatient department. In this section, we develop an empirical model with the limited objective of demonstrating how the qualitative findings of our queueing model can be validated using real data.

5.1. Modeling Approach

As discussed earlier, ambulance diversion is one of several options available to EDs to mitigate overcrowding. Other options include creating temporary surge capacity by placing beds and stretchers in hallways and early discharge for inpatients with stable condition. Some hospitals might adopt

a strategic decision to accept all incoming ambulances irrespective of the extent of overcrowding (GAO 2003) due either to their location (e.g. rural) or their mission (e.g. community hospitals). In other words, it is possible that EDs self-select themselves into the sub-sample that has positive diversion hours. Hence estimating the extent of diversion only for this sub-sample would lead to biased coefficient estimates for the entire population. We mitigate this problem by endogenizing EDs' decision to use ambulance diversion and estimating it jointly with the extent of diversion using a selection model (Amemiya 1984, Greene 2008) shown below:

$$y_{1i} = \alpha' \mathbf{Z}_i + \varepsilon_{1i} \quad (4a)$$

$$y_{2i} = \begin{cases} \gamma' \mathbf{X}_i + \varepsilon_{2i} & \text{if } y_{1i} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (4b)$$

(4a), referred to as the choice equation, governs whether the EDs choose to employ ambulance diversion ($y_1 > 0$) or not ($y_1 \leq 0$). In the former case, (4b), referred to as the level equation, determines the extent of diversion. The independent variables X and Z in the two equations might be the same or different. The error terms ε_1 and ε_2 are assumed to be distributed according to bivariate normal distribution. We assume that only the sign of the selection variable y_1 can be inferred but not its magnitude. Hence the selection equation (4a) is reformulated as follows by introducing another variable w_{1i} where $w_{1i} = 1$ if $y_{1i} > 0$ and $w_{1i} = 0$ otherwise:

$$\mathbb{P}(w_{1i} = 1 | \mathbf{Z}_{1i}) = \Phi(\alpha' \mathbf{Z}_i) \quad (5)$$

where $\Phi(\cdot)$ is the cumulative normal distribution.

Two methods are commonly used in the econometrics literature to estimate the selection model (4b) and (5). The first method, called Full Information Maximum Likelihood (FIML) is exact and involves maximizing the complete likelihood function involving the double integral over the bivariate normal distribution of error terms. The second method, known as Limited Information Maximum Likelihood estimation (LIML), explicitly recognizes the problem of estimating equation (4b) based on the sub-sample as that of an omitted variable bias (Heckman 1979) and involves a two-step procedure. The first step involves estimating (5) using a Probit model and calculating

$\hat{\lambda}_i = \frac{\phi(\alpha' \mathbf{z}_i)}{\Phi(\alpha' \mathbf{z}_i)}$ known as the Inverse Mills Ratio (IMR). In the second step, IMR is introduced in the level equation (4b), which is then estimated using OLS on the sub-sample.

While the first method is more exact, the likelihood function is not guaranteed to be concave and maximizing it might also present some computational difficulties. It is also known to be sensitive to deviation from normality of error terms and measurement errors in the dependent variable (Stapleton and Young 1984). The second method, while being approximate, is easier to compute and also more robust to misspecifications (Leung and Yu 2000). There is some evidence that it can also be more efficient than the first method for small samples (Leung and Yu 1996). Hence, we use both methods to estimate our model and compare their results.

5.2. Data

We study ambulance diversion in hospitals licensed in the state of California. In accordance with the state law, all licensed hospitals report their financial and operational data to the Office of the Statewide Hospital Planning and Development (OHSPD) using an online system known as ALIRTS (Automated Licensing Information and Report Tracking System). The data is then made publicly available after minimal input quality control edits.

We use two separate datasets containing capacity and utilization data available from from the OSHPD website (<http://www.oshpd.ca.gov/HID/DataFlow/HospData.html>) for calendar year 2004. First dataset called State Utilization Data File for Hospitals contains basic licensing information (ownership, location and type of the facility) and annual utilization information (licensed bed capacity and patient census for different bed types, number of visits to the ED and total number of diversion hours) for 491 hospitals. Second dataset called Hospital Annual Financial Data File contains financial data (net patient revenue and costs by payer type, balance sheet, staff productivity) in addition to some licensing and utilization data for 451 hospitals. After merging the two datasets and retaining only the records common to both datasets, we end up with 431 hospitals. Since our objective is to study the extent of ambulance diversion in emergency rooms, we drop the hospitals which either did not have an ED (zero ED beds reported) or did not operate the ED

for the period under consideration (zero ED visits reported). This results in 318 hospitals being retained for analysis.

5.3. Measures

In Section 4, we show that the outcome variable of interest, the fraction of time on diversion δ , is decreasing in the (normalized) spare capacity of the inpatient department $\beta_2 = (1 - \rho_2)\sqrt{N_2}$ and decreasing in the (normalized) size of the ED $\frac{N_1}{\sqrt{N_2}}$. Thus in order to estimate the model empirically, we need to construct measures for the underlying independent variables ρ_2, N_2 and N_1 , where N_2 represents the number of inpatient beds that are relevant to the flow of patient admissions from the ED and ρ_2 represents the utilization or occupancy of these beds.

Our dataset contains the number of licensed inpatient beds for the following categories: general acute care (GAC), acute psychiatric, skilled nursing and intermediate care beds. GAC is further subclassified into medical/surgical, pediatric, perinatal, intensive care, coronary care, acute respiratory, burn, rehabilitation, and neonatal intensive care beds. A number of empirical studies have highlighted the impact of ICU beds on the extent of ambulance diversion in the ED (McConnell et al. 2005). Similarly, past surveys have also highlighted the lack of ICU beds as one of the most important reasons for going on ambulance diversion (GAO 2003, McManus 2001). Based on these studies, we choose the number of intensive care unit beds to construct a measure for N_2 . However, licensed beds typically do not reflect the true bed capacity since hospitals might staff only a fraction of all the licensed beds depending on the patient load (Green 2002). Data on staffed beds is available only in aggregate whereas data on licensed beds is available for all the subcategories mentioned above. Hence, we calculate the number of staffed ICU beds (N_I)³ and use it as a measure of N_2 . Similarly, we calculate the theoretical utilization of staffed ICU beds ρ_I using the actual census of ICU patients reported and the staffed ICU beds N_I calculated above. We then use these measures to construct variables of our interest as follows: $\text{KAPPA} = \frac{N_1}{\sqrt{N_I}}$, which reflects the size of the ED relative to the ICU and $\text{BETA2} = (1 - \rho_I)\sqrt{N_I}$ which reflects the spare capacity in the

³ staffed icu beds = $\frac{\text{licensed ICU beds}}{\text{total licensed beds}} * \text{total staffed beds}$

ICU normalized for the size of the ICU. We use the full names instead of their symbols to highlight the difference between the proxy measures and the variables. Since the observation period for all hospitals is the same (calendar year 2004), we use total hours on diversion (DIV_HOURS) as a proxy for the fraction of time on diversion. In addition, we use a dummy variable to denote whether the hospital had any positive diversion hours (AMB_DIV=1 if DIV_HOURS>0).

5.4. Controls

The independent variables described above attempt to explain the extent of ambulance diversion conditional on the ED's decision to use ambulance diversion to mitigate overcrowding. However, this decision of whether to employ ambulance diversion or not might itself depend on other characteristics of the hospital such as location and ownership structure. For example, many hospitals in rural areas might decide not to go on ambulance diversion since there are no alternate hospitals in the vicinity. Similarly, it is plausible that nonprofit or academic hospitals are less likely to divert ambulances because of their mission. Also, trauma centers might adopt different ambulance diversion policies than EDs which are not designated as trauma centers. Hence, we constructed measures for these control variables.

In our data, hospitals are designated as rural hospitals based on Section 124840 of the California Health and Safety Code which includes the following criteria: acute care hospital with less than 76 beds and located in a census dwelling place with less than 15000 residents as per the 1980 census. We create a dummy variable to indicate the location of the hospital (RURAL=1 for rural and RURAL=0 for urban). We also classify ownership control of the hospitals into the following categories: GOVERNMENT (City and/or County, District), NONPROFIT (Not-for-profit including university hospitals), and INVESTOR (for-profit concerns including partnerships, and public and private companies). The California EMS Authority assigns one of four designations to each ED with a trauma center depending on its capabilities to treat and stabilize trauma patients with level 1 being the most capable and level 4 being the least capable. We create a dummy variable to indicate if an ED was designated as a trauma center in one of these categories (TRAUMA=1) or not (TRAUMA=0).

5.5. Descriptive Statistics

Since we use the average occupancy of ICU beds as a measure of the inpatient utilization, we dropped hospitals that do not have an ICU (zero ICU licensed beds reported) retaining 295 hospitals for our analysis. Table 3 compares the sample of hospitals having an ICU and those without an ICU on a few key measures. Only one of the 23 hospitals without an ICU experienced ambulance diversion (3 hours) indicating that ambulance diversion is a much bigger problem in hospitals with ICU. On average, the hospitals without ICUs are smaller than those with ICUs and had smaller EDs.

Table 3 Comparison of measures in hospitals with and without ICUs.

Measure	Hospitals without ICUs (N=23)	Hospitals with ICUs (N=295)	p-value
Diversion Hours	0.13 (0.63)	790.17 (1261.62)	< 0.01
Staffed inpatient beds	42.43 (31.54)	204.50 (143.37)	< 0.01
ED beds	4.26 (2.40)	17.95 (11.08)	< 0.01

Note: Each cell contains the sample mean with the standard error in parentheses

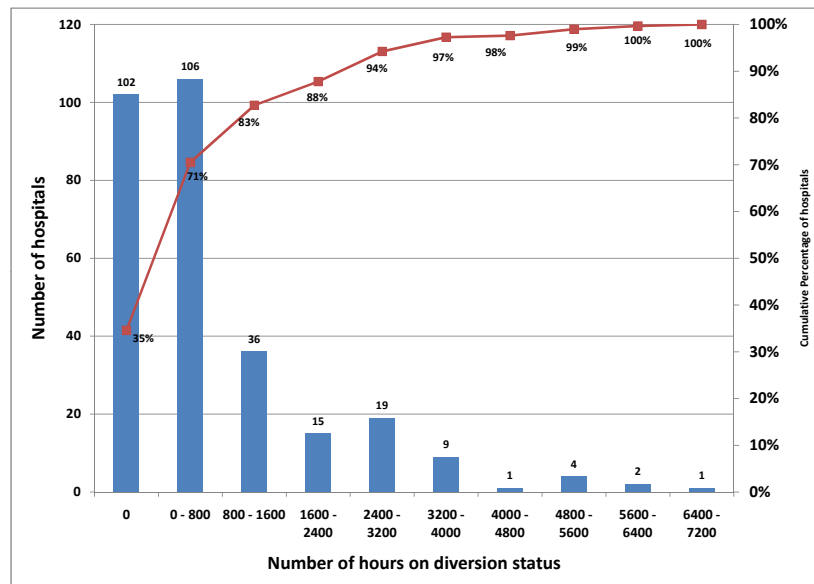
Table 4 presents the descriptive statistics for the basic and the standardized continuous variables and our categorical control variables. We find a huge variation in DIV_HOURS (from zero to 7170 hours); the distribution is shown in Figure 3. Approximately 35% of the EDs had no diversion hours, again suggesting that some hospitals make a strategic decision of not going on diversion as discussed above. Moreover, the long tail and the fact that standard deviation is much larger than the mean indicates that the data is not normally distributed. Hence we transformed our dependent variable by taking its natural logarithm. Note that the utilization of the ICU beds exceeds 100% in some cases. This is possible because ICU patients are occasionally placed in other wards if ICU beds are full.

5.6. Estimation and Results

We use AMB_DIV to denote the choice of hospitals in (5) and DIV_HOURS to denote the extent of diversion in (4b). In accordance with the queueing model we include variables BETA2, KAPPA and

Table 4 Descriptive Statistics.

Variable Name	Mean	Std. dev.	Min.	Max.
DIV_HOURS	790.17	1261.62	0.00	7170.00
N_1	17.95	11.08	1.00	64.00
N_I	17.09	17.42	2.00	189.00
ρ_I	0.66	0.31	0.00	2.05
KAPPA	4.65	1.94	0.29	11.63
BETA2	1.30	1.31	-2.99	6.65
RURAL	0.13			
INVESTOR	0.60			
GOVERNMENT	0.16			
TRAUMA	0.18			

Figure 3 Distribution of ambulance diversion hours in study sample

their interaction term $BETA2 * KAPPA$ in (4b). However, since the decision to choose ambulance diversion can be influenced by structural variables (location, ownership and trauma center designation) in addition to the operational variables, we also include these (RURAL, GOVERNMENT, INVESTOR, TRAUMA) in (5). Due to the presence of the interaction term $KAPPA * BETA2$, we

center the variables KAPPA and BETA2 around their respective means (\overline{KAPPA} and $\overline{BETA2}$) in both equations to reduce nonessential multicollinearity and to provide easy interpretation of the coefficients (Cohen et al. 2003). The results for this model (MODEL I) are shown in the first two columns of Table 5. The first column represents the LIML estimation and the second column represents the FIML estimation method.

First, note that the coefficient of the Inverse Mills Ratio (IMR) is statistically significant in the LIML estimation. This rejects the null hypothesis that the sub-sample of hospitals with positive diversion hours is randomly selected from the larger sample (Leung and Yu 1996, Melino 1982). Thus, we find evidence that some hospitals make explicit decisions to accept all ambulances irrespective of the extent of overcrowding. Examining the coefficient estimates for the “Choice Equation” we find that rural hospitals are less likely to use ambulance diversion whereas trauma centers and private hospitals are more likely to use ambulance diversion. Many rural hospitals might be forced to accept all ambulances due to the lack of other hospitals in their catchment area. On the other hand, diverting ambulances with trauma patients might be more effective in ensuring prompt care than accepting them in an overcrowded ED. Interestingly, we also found that the hospitals whose relative size of the ED compared to the ICU is larger are more likely to choose ambulance diversion. This suggests a mechanism by which hospitals with larger EDs attract more patients that need hospitalization and place a higher burden on their ICUs, thereby congesting them and resulting in larger durations of ambulance diversion status. However, we could not directly test this mechanism since we did not have data on the percent of emergency visits admitted to ICU.

For the “Level Equation”, the interaction term KAPPA*BETA2 is significant at 10% level indicating that KAPPA and BETA2 moderate each other’s effect on the extent of diversion, conditional on the hospital’s decision to use ambulance diversion. This is in conformance with the analytical expression for the fraction of time on diversion (2) which includes a product term $\kappa\beta_2$. The coefficients of BETA2 and KAPPA are not significant in Table 5. However, in the presence of interaction, these coefficients represent the conditional effect of these variables at the mean values

Table 5 Estimation results for selection models.

	MODEL I		MODEL II	
	LIML	FIML	LIML	FIML
LEVEL EQUATION				
INTERCEPT	6.54*** (0.29)	7.33*** (0.16)	6.69*** (0.3)	7.35*** (0.15)
BETA2 - $\overline{BETA2}$	-0.06 (0.09)	0.01 (0.10)	-0.08 (0.08)	-0.04 (0.05)
KAPPA - \overline{KAPPA}	-0.07 (0.06)	-0.16** (0.04)	-0.01 (0.06)	-0.08*** (0.03)
(KAPPA - \overline{KAPPA})*(BETA2 - $\overline{BETA2}$)	-0.06* (0.04)	-0.06 (0.04)	-0.05 (0.04)	-0.04* (0.02)
INVERSE MILLS RATIO	-0.98* (0.53)		-1.28** (0.54)	
CHOICE EQUATION				
INTERCEPT	0.46*** (0.11)	0.55*** (0.08)	0.47*** (0.11)	0.47*** (0.09)
BETA2 - $\overline{BETA2}$	-0.04 (0.06)	-0.01 (0.04)		
KAPPA - \overline{KAPPA}	0.13*** (0.04)	0.08*** (0.03)		
(KAPPA - \overline{KAPPA})*(BETA2 - $\overline{BETA2}$)	0.02 (0.03)	0.03 (0.02)		
RURAL	-1.28*** (0.27)	-0.48*** (0.04)	-1.37*** (0.26)	-0.39*** (0.04)
TRAUMA	0.38* (0.23)	0.11 (0.08)	0.36* (0.22)	0.15 (0.18)
INVESTOR	0.68*** (0.22)	-0.24*** (0.04)	0.5** (0.21)	-0.15*** (0.03)
GOVERNMENT	-0.46 (0.3)	-0.28*** (0.03)	-0.47* (0.29)	-0.27*** (0.02)
Log Likelihood	-520.62	-517.19	-161.92	-154.55

Notes: Standard Errors are shown in parentheses. Log Likelihood for the LIML estimation method is corresponding to the Probit model for the selection equation. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

of each other. In order to get a more complete picture, we also estimate the conditional effect of each of these variables at different values of the other variable, specifically two standard deviations above and below the mean. These results are shown in Table 6. We find that coefficient of BETA2

is significant at 10% level for very high values of KAPPA and coefficient of KAPPA is significant at 10% for high values of BETA2 thus providing support for our theoretical results (Proposition 1 and 3).

Table 6 Significance of simple slopes for KAPPA and BETA2 in MODEL I.

	SIMPLE SLOPE OF BETA2		SIMPLE SLOPE OF KAPPA	
	LIML	FIML	LIML	FIML
KAPPA - $\overline{KAPPA} = -2*\sigma_{KAPPA}$	0.22 (0.19)	0.29 (0.22)		
KAPPA - $\overline{KAPPA} = 0$	-0.06 (0.09)	0.01 (0.10)		
KAPPA - $\overline{KAPPA} = 2*\sigma_{KAPPA}$	-0.34* (0.20)	-0.27 (0.24)		
BETA2 - $\overline{BETA2} = -2*\sigma_{BETA2}$			0.12 (0.12)	0.03 (0.14)
BETA2 - $\overline{BETA2} = 0$			-0.07 (0.06)	-0.16** (0.07)
BETA2 - $\overline{BETA2} = 2*\sigma_{BETA2}$			-0.26* (0.15)	-0.35** (0.17)

Note: Standard Errors are shown in parentheses. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$

This provides an interesting interpretation about the bottleneck in the patient flow process. The ICU beds are the bottleneck when the size of the ED is relatively large compared to the ICU (higher values of KAPPA): the congestion in the ED and the consequent ambulance diversion is increasing in the utilization of the ICU beds. On the other hand, the ED is the bottleneck when the utilization of the ICU normalized for its size is very low (high values of BETA2): the extent of ambulance diversion is decreasing in the number of ICU beds. This partly explains the findings from previous single-hospital studies that find ICU and other inpatient departments to be the primary driver for ambulance diversion as most of these studies are conducted in large academic hospitals, which are likely to have relatively larger EDs.

The results for the FIML estimation are quite similar. Rural hospitals are less likely to adopt ambulance diversion compared to urban hospitals; private and Government hospitals are less likely to adopt ambulance diversion compared to non-profit hospitals. On the other hand, hospitals with relatively large EDs compared to the ICU and trauma centers are more likely to employ ambulance diversion though the coefficient for the latter is not statistically significant. The coefficient of the

interaction term in the second column of Table 5 is very close to that in the LIML estimation but not significant. Examining the coefficients in the second and fourth column of Table 6, we again observe the phenomenon of shifting bottlenecks described earlier except that the coefficient of BETA2 was negative but not significant for large values of KAPPA.

We also estimate a variant (MODEL II) of the above model specification wherein we retained only the structural variables (RURAL, GOVERNMENT, INVESTOR, TRAUMA) in the selection equation. The results for this model are shown in columns 3 and 4 of Table 5. The results for both FIML and LIML estimation remain roughly similar except that the interaction term KAPPA*BETA2 is significant at 10% level in the FIML estimation and loses significance in the LIML estimation. Since MODEL II is nested in MODEL I, we use the likelihood ratio test to test the hypothesis that the coefficients of KAPPA, BETA2 and KAPPA*BETA2 are jointly zero. We reject this hypothesis for both FIML ($p < 0.1$) and LIML ($p < 0.01$) estimation methods indicating that the earlier model provides a better fit of the data.

5.7. Limitations

Our empirical analysis has several limitations stemming partly from the limitations of our data and partly from the complexities of the queueing model. Our study sample is not drawn randomly from the population of hospitals in the U.S. and hence our results cannot be directly generalized beyond California. While it is reasonable to expect that similar operational factors impact ambulance diversion in other states as well, California differs from many other states in certain important respects.

Unlike California, many states require hospitals to obtain an approval (called Certificate of Need) from the local health planning agencies prior to expanding their capacity (Cimasi 2005). To the extent that hospitals face such obstacles to differing degrees in the ED and the inpatient departments, we might expect to see a different impact of capacity of these two components on extent of diversion. California is also unique in that its “Medicaid Disproportionate Share Hospital” program provides vital funds to hospitals providing care to the most vulnerable populations (Melnick et al.

2004). This might reduce the gap in resources and, consequently, capacity among hospitals that might be more pronounced in other states. It is also likely that local regulations governing the validity and appropriateness of ambulance diversion might differ across states.

Our data is aggregate and represents annual averages for inpatient occupancy and staffed beds and year end number for licensed beds. It has been widely observed that demand for emergency services follows a cyclic pattern with typically more visits on the weekdays than the weekends and more arrivals in mornings than afternoons (Green et al. 2006, Burt and McCaig 2006). Inpatient admission and discharge processes also follow similar cyclic patterns (McManus et al. 2003). To the extent that the number of staffed beds cannot be continually adjusted to match these patterns, our aggregate data would have underestimated the magnitude of undercapacity and overestimated the magnitude of overcapacity. This creates a bias against finding significant effect on inpatient occupancy on ED overcrowding and ambulance diversion.

Even the stylized queueing network model of patient flow from the ED to the inpatient department is quite complicated and requires a series of approximations to derive analytical results. Moreover, these analytical expressions are highly nonlinear in the parameters that need to be estimated. This prevents us from undertaking a structural estimation of the parameters of our queueing model. Rather, the empirical results from the analysis of our linear model demonstrate the qualitative impact of inpatient and ED capacity on ambulance diversion.

6. Conclusion

In this paper we present theoretical as well as empirical examination of the phenomenon of ED overcrowding and ambulance diversion in hospitals. Ambulance diversion is one of the most serious problems facing the emergency health care systems in the US and many other developed countries. While earlier investigations have shown that higher occupancy in the inpatient departments such as the ICU is associated with greater extent of ambulance diversion in the ED, these studies have lacked the theoretical framework to compare the effect of hospital size on the strength of this association.

In contrast, we employ a queueing network model for the flow of patients between the ED and the inpatient department of the hospital and analyze it in the heavy traffic regime. Our analysis shows that the extent of ambulance diversion is increasing in the utilization of the inpatient department and decreasing in the size of the ED, both appropriately normalized by the size of the hospital. Since these theoretical predictions explicitly account for the size of the hospital, we are able to test them on a cross-sectional data of hospitals in California. The results of our empirical analysis provides partial support for the theoretical findings. These empirical findings need to be confirmed by using data from other states and by using data that is less aggregated.

We also find that the capacity of the inpatient department and the ED interact with each other in determining their impact on ambulance diversion, an effect that has not yet been identified in the literature. Specifically, inpatient department tends to be the bottleneck in hospitals where the ED is large relative to the hospital. Conversely, ED tends to be the bottleneck in hospitals where the inpatient department is relatively sparsely occupied.

This result has important implications for the formulation of public policy. It suggests that different measures might be required to mitigate ambulance diversion in hospitals with different structural characteristics rather than a single, across the board prescription such as reduction of inpatient utilization. Our findings also suggest that hospitals need to determine the size of their ED and inpatient departments concurrently to avoid a mismatch of capacity in the two components leading to ED overcrowding and ambulance diversion. Future work could explore the development of detailed models for planning bed capacity in hospitals.

Appendix

A. Estimation on Delay Probability P_d

In this section, we derive the function $\tilde{P}_d(N_1, N_2, \beta_2, \kappa)$ that we use to approximate the delay probability P_d .

Let ρ_1 denote the traffic intensity of station 1 (the ED). It is clear that

$$\rho_1 = (\lambda_w + (1 - \delta)\lambda_a)m_1/(N_1 - B). \quad (6)$$

Using Proposition 1 of Halfin and Whitt (1981), we can approximate the delay probability in the ED by $\left[1 + \frac{\beta_1 \Phi(\beta_1)}{\phi(\beta_1)}\right]^{-1}$, where

$$\beta_1 = \sqrt{N_1 - B}(1 - \rho_1). \quad (7)$$

Therefore, to obtain an estimate of P_d , we need estimates of B , the average number of boarding patients, and ρ_1 , the traffic intensity of the ED.

We first approximate B using the steady state distribution of \tilde{Q}_2 given in Whitt (2004),

$$\tilde{B}(N_2, \beta_2, \kappa) = \sqrt{N_2} \frac{1 - e^{-\kappa\beta_2}(1 + \kappa)}{\beta_2(1 - e^{-\kappa\beta_2} + \beta_2 \frac{\phi(\beta_2)}{\Phi(\beta_2)})} \quad (8)$$

Next, from (6), we can approximate ρ_1 as

$$\tilde{\rho}_1(N_1, N_2, \beta_2, \kappa) = \frac{\lambda_1(1 - \tilde{\delta}(N_2, \beta_2, \kappa)) + \lambda_2}{N_1 - \tilde{B}(N_2, \beta_2, \kappa)} m_1, \quad (9)$$

where $\tilde{\delta}$ and \tilde{B} are given in (2) and (8) respectively. We can then substitute (9) and (8) in (7) to obtain $\tilde{\beta}_1(N_1, N_2, \beta_2, \kappa)$. And $\tilde{P}_d(N_1, N_2, \beta_2, \kappa) = \left[1 + \frac{\tilde{\beta}_1(N_1, N_2, \beta_2, \kappa) \Phi(\tilde{\beta}_1(N_1, N_2, \beta_2, \kappa))}{\phi(\tilde{\beta}_1(N_1, N_2, \beta_2, \kappa))}\right]^{-1}$.

B. Proofs

Proof of Proposition 1 To prove result (i), let $g_1(\beta_2) = (\sqrt{N_2} - \beta_2)e^{\kappa\beta_2/2}$ and $g_2(\beta_2) = (e^{\kappa\beta_2/2} - e^{-\kappa\beta_2/2})/\beta_2 + e^{\kappa\beta_2/2}\Phi(\beta_2)/\phi(\beta_2)$. Since $\tilde{\delta}(N_2, \beta_2, \kappa) = [g_1(\beta_2)g_2(\beta_2)]^{-1}$, it is sufficient to show that g_1 is nondecreasing and g_2 is increasing in β_2 . To show g_1 is nondecreasing, we look at the first order derivative:

$$g_1'(\beta_2) = e^{\kappa\beta_2/2}[\kappa(\sqrt{N_2} - \beta_2)/2 - 1] \quad (10)$$

$$= e^{\kappa\beta_2/2}(K\rho_2/2 - 1) \geq 0. \quad (11)$$

For g_2 , it is easy to see that $e^{\kappa\beta_2/2}\Phi(\beta_2)/\phi(\beta_2)$ is increasing in β_2 . Therefore, it remains to show that $g_3(\beta_2) = (e^{\kappa\beta_2/2} - e^{-\kappa\beta_2/2})/\beta_2$ is nondecreasing in β_2 . Again, we take the first order derivative:

$$g_3'(\beta_2) = [(\kappa\beta_2/2)(e^{\kappa\beta_2/2} + e^{-\kappa\beta_2/2}) - e^{\kappa\beta_2/2} + e^{-\kappa\beta_2/2}]/\beta_2^2 = f(\kappa\beta_2/2)/\beta_2^2,$$

where $f(x) = x(e^x + e^{-x}) - e^x + e^{-x}$. Obviously $f(0) = 0$. Moreover $f(x)$ is nondecreasing in x because $f'(x) = x(e^x - e^{-x}) \geq 0$. Therefore, $f(x) \geq 0$ for $x \geq 0$, proving $g_3'(\beta_2) \geq 0$.

To prove result (ii), we can rewrite (2) as

$$\tilde{\delta}(N_2, \beta_2, \kappa) = \frac{\beta_2}{(\sqrt{N_2} - \beta_2) \left(e^{\kappa\beta_2} - 1 + e^{\kappa\beta_2} \beta_2 \frac{\Phi(\beta_2)}{\phi(\beta_2)} \right)}.$$

The result then follows because the function $\left(e^{\kappa\beta_2} - 1 + e^{\kappa\beta_2} \beta_2 \frac{\Phi(\beta_2)}{\phi(\beta_2)} \right)$ is increasing in κ .

Proof of Proposition 2 We can rewrite equation (8) as

$$\tilde{B}(N_2, \beta_2, \kappa) = \sqrt{N_2} \left(\frac{1}{\beta_2} - \frac{\kappa e^{-\kappa\beta_2}}{(1 - e^{-\kappa\beta_2})} \right) / \left(1 + \frac{\beta_2 \Phi(\beta_2)}{\phi(\beta_2)(1 - e^{-\kappa\beta_2})} \right).$$

It is easy to see that \tilde{B} is increasing in κ because $\frac{\kappa e^{-\kappa\beta_2}}{1 - e^{-\kappa\beta_2}} = \kappa / (e^{\kappa\beta_2} - 1)$ is decreasing in κ . Since $\tilde{\delta}$ is decreasing in κ , $\tilde{\beta}_1$ is decreasing in κ . Therefore, \tilde{P}_d is increasing in κ .

Proof of Proposition 3 It is equivalent to show that κ^* is increasing in N_1 . It is easy to see that \tilde{P}_d is decreasing in N_1 because, clearly, $\tilde{\beta}_1$ is increasing in N_1 . Now for any $N_{1,1} < N_{1,2}$, let $\kappa_1 = \kappa^*(N_{1,1}, N_2, \beta_2, \bar{P}_d)$ and $\kappa_2 = \kappa^*(N_{1,2}, N_2, \beta_2, \bar{P}_d)$. It suffices to show $\kappa_1 < \kappa_2$.

Suppose $\kappa_1 \geq \kappa_2$. Because $N_{1,1} < N_{1,2}$, we have $\tilde{P}_d(N_{1,1}, N_2, \beta_2, \bar{P}_d) > \tilde{P}_d(N_{1,2}, N_2, \beta_2, \bar{P}_d)$, which contradicts with the fact that $\tilde{P}_d(N_{1,1}, N_2, \beta_2, \bar{P}_d) = \tilde{P}_d(N_{1,2}, N_2, \beta_2, \bar{P}_d) = \bar{P}_d$.

References

- J. Adams. Personal communication, 2008.
- T. Amemiya. Tobit models: A survey. *Journal of Econometrics*, 24(1-2):3–61, 1984.
- American Hospital Association. *Trend Watch Chartbook 2005*. Online. Available: <http://www.ahapolicyforum.org/ahapolicyforum/trendwatch/chartbook2005.htm>, 2005.
- B. R. Asplin, D. J. Magid, K. V. Rhodes, L. I. Soldberg, N. Lurie, and C. A. Carmago. A conceptual model of emergency department overcrowding. *Annals of Emergency Medicine*, 42(2):173–180, 2003.
- C. W. Burt and L. F. McCaig. Staffing, capacity, and ambulance diversion in emergency departments: United states, 2003–04. *CDC Advance data from vital and health statistics*, 376, 2006.
- C. W. Burt, R. McCaig, and R. Valverde. Analysis of ambulance transports and diversions among us emergency departments. *Annals of Emergency Medicine*, 47(4):317–326, 2006.
- R. J. Cimasi. *U.S. healthcare certificate of need sourcebook*. Frederick, MD: Beard Books, 2005.
- J. Cohen, P. Cohen, S. G. West, and L. S. Aiken. *Applied multiple regression / correlation analysis for the behavioral sciences (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 2003.
- J. K. Cooper and T. M. Corcoran. Estimating bed needs by means of queuing theory. *The New England Journal of Medicine*, 291, 1974.
- R. W. Derlet and J. R. Richards. Overcrowding in the nation’s emergency departments: Complex causes and disturbing effects. *Annals of Emergency Medicine*, 35(1):63–68, 2000.
- R. W. Derlet, J. R. Richards, and R. L. Kravitz. Frequent overcrowding in u.s. emergency departments. *Academic Emergency Medicine*, 8(2):151–155, 2001.
- M. Eckstein and L. S. Chan. The effect of emergency department crowding on paramedic ambulance availability. *Annals of Emergency Medicine*, 43(1):100–105, 2004.
- J. A. Fitzsimmons. A methodology for emergency ambulance deployment. *Management Science*, 19(6):627–636, 1973.

- A. J. Forster, I. Stiell, G. Wells, A. J. Lee, and C. Van Walraven. The effect of hospital occupancy on emergency department length of stay and patient disposition. *Academic Emergency Medicine*, 10(2):127–133, 2003.
- GAO. *Hospital emergency departments: Crowded conditions vary among hospitals and communities*. Washington DC: U.S. General Accounting Office, 2003.
- Y. Gerchak, D. Gupta, and M. Henig. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science*, 42(3):321–334, 1996.
- P. G. Gosselin. Amid nationwide prosperity, ers see a growing emergency, Aug 6 2001.
- L. V. Green. How many hospital beds? *Inquiry*, 39:400–412, 2002.
- L. V. Green and P. J. Kolesar. Improving emergency responsiveness with management science. *Management Science*, 50(8):1001–10014, 2004.
- L. V. Green and V. Nguyen. Strategies for cutting hospital beds: The impact on patient service. *Health Services Research*, 36(2):421–442, 2001.
- L. V. Green, J. Soares, J. F. Giglio, and R. A. Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, 2006.
- W. H. Greene. *Econometric analysis (6th ed.)*. Upper Saddle River, NJ: Pearson Education, Inc., 2008.
- S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.
- J. H. Han, C. Zhou, D. J. France, S. Zhong, I. Jones, A. B. Storrow, and D. Aronsky. The effect of emergency department expansion on emergency department overcrowding. *Academic Emergency Medicine*, 14(4):338–343, 2007.
- J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.
- J. D. Hershey, E. N. Weiss, and M. A. Cohen. A stochastic service network model with application to hospital facilities. *Operations Research*, 29(1):1–22, 1981.
- P. L. Joskow. The effects of competition and regulation on hospital bed supply and the reservation quality of the hospital. *The Bell Journal of Economics*, 11(2):421–447, 1980.
- J. Kennedy, K. Rhodes, C. A. Walls, and B. Asplin. Access to emergency care: Restricted by long waiting times and cost and coverage concerns. *Annals of Emergency Medicine*, 43(5):567–573, 2004.
- S. F. Leung and S. Yu. Collinearity and two-step estimation of sample selection models: Problems, origins, and remedies. *Computational Economics*, 15(3):173–199, 2000.
- S. F. Leung and S. Yu. On the choice between sample selection and two-part models. *Journal of Econometrics*, 72(1-2):197–229, 1996.

- K. J. McConnell, C. F. Richards, M. Daya, S. H. Bernell, C. C. Weathers, and R. A. Lowe. Effect of increased icu capacity on emergency department length of stay and ambulance diversion. *Annals of Emergency Medicine*, 45(5):471–478, 2005.
- M. McManus. *Emergency department overcrowding in Massachusetts: Making room in our hospitals*. Waltham, MA: The Massachusetts Health Policy Forum, 2001.
- M. L. McManus, M. C. Long, A. Cooper, J. Mandell, D. M. Berwick, M. Pagano, and E. Litvak. Variability in surgical caseload and access to intensive care services. *Anesthesiology*, 98(6):1491–1496, 2003.
- A. Melino. Testing for sample selection bias. *The Review of Economic Studies*, 49(1):151–153, 1982.
- G. A. Melnick, A. C. Nawathe, A. Bamezai, and L. Green. Emergency department capacity and access in california, 1990–2001: An economic analysis. *Health Affairs.*, 23(3), 2004.
- C. T. Merrill and A. Elixhauser. *Hospitalization in the United States, 2002: HCUP Fact Book No. 6*. Rockville, MD: Agency for Healthcare Research and Quality, 2005.
- J. G. Mulligan. The stochastic determinants of hospital-bed supply. *Journal of Health Economics*, 4(2): 177–181, 1985.
- K. Ramdas and J. Williams. An empirical investigation into the tradeoffs that impact on-time performance in the airline industry. *Working Paper*, 2008.
- E. S. Savas. Simulation and cost-effectiveness analysis of new york’s emergency ambulance service. *Management Science*, 15(12):B608–B627, 1969.
- M. J. Schull, K. Lazier, M. Vermeulen, S. Mawhinney, and L. J. Morrison. Emergency department contributors to ambulance diversion: A quantitative analysis. *Annals of Emergency Medicine*, 41(4):467–476, 2003a.
- M. J. Schull, L. J. Morrison, M. Vermeulen, and D. A. Redelmeier. Emergency department gridlock and out-of-hospital delays for cardiac patients. *Academic Emergency Medicine*, 10(7):709–716, 2003b.
- M. J. Schull, M. Vermuelen, G. Slaughter, L. Morrison, and P. Daly. Emergency department crowding and thrombolysis delays in acute myocardial infarction. *Annals of Emergency Medicine*, 44(6):577–585, 2004.
- N. Shute and M. B. Marcus. Crisis in the er, Sep 10 2001.
- L. T. Soldberg, B. R. Asplin, R. M. Weinick, and D. J. Magid. Emergency department crowding: Consensus development of potential measures. *Annals of Emergency Medicine*, 42(6):824–834, 2003.
- D. Stapleton and D. Young. Censored normal regression with measurement error on the dependent variable. *Econometrica*, 52(3):737–760, 1984.
- C. Swoveland, D. Uyeno, I. Vertinsky, and R. Vickson. Ambulance locations: A probabilistic enumeration approach. *Management Science*, 20(4):686–698, 1973.

- The Lewin Group. *Emergency department overload: A growing crisis - The results of the AHA survey of emergency department (ED) and hospital capacity*. Falls Church, VA: American Hospital Association, 2002.
- G. Vassilopoulos. A simulation model for bed allocation to hospital inpatient departments. *Simulation*, 45(5):233–241, 1985a.
- G. Vassilopoulos. Allocating doctors to shifts in an accident and emergency department. *Journal of the Operational Research Society*, 36(6):517–523, 1985b.
- W. Whitt. A diffusion approximation for the G/GI/n/m queue. *Operations Research*, 52(6):922–941, 2004.