

# 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”. The objective of this paper is 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 propose a simple queueing model to describe the patient flow between the ED and the inpatient department. We analyze this model using two different approximations - heavy traffic and fluid - to derive two separate sets of measures for inpatient occupancy and ED size. We use these sets of measures to form hypotheses and test them by estimating a sample selection model using data on a cross-section of hospitals from California. We find that the measures derived from the heavy traffic approximation provide better explanation of the data than those derived from the fluid approximation. For the former specification, we find that the fraction of time that the ED spends 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. In addition, controlling for these hospital-specific factors, we find that the fraction of time on diversion at a hospital increases with number of hospitals in its neighborhood. We also find that certain hospitals, owing to their location, ownership and trauma center status, 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 either to qualitative commentary and surveys (Derlet and Richards 2000, The Lewin Group 2002, GAO 2003, Burt and McCaig 2006) or to single hospital empirical studies (Schull et al. 2003a, Han et al. 2007, McConnell et al. 2005, Forster et al. 2003).

The objective of this paper is to inform this discussion by conducting a cross-sectional analysis about the impact of various operational characteristics of the hospital on the extent of diversion. Since the hospitals in the cross-section display considerable heterogeneity with respect to size, occupancy, ownership, location etc., we find that raw measures of hospital size and occupancy are not sufficient to explain the variation in the extent of diversion in our sample. Hence we make two significant improvements over this basic approach: (i) we employ analytical models to develop appropriate measures for our empirical study, and (ii) we endogenize hospitals' decision to use diversion to mitigate overcrowding. We briefly describe our approach below.

We develop a two-station queueing model <sup>1</sup> 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 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 threshold<sup>2</sup>. Since our objective is to

<sup>1</sup> See Hershey et al. (1981) and Cooper and Corcoran (1974) for early examples of application of queueing theory to hospitals

<sup>2</sup> Such policies are routinely followed in practice (Green 2002, Adams 2008) and are discussed in greater detail in Section 4.1.2.

derive appropriate analytical measures of inpatient capacity and ED size, which we can employ in our empirical study, we consider two approximations that are appropriate for large and busy systems - heavy traffic and fluid. We derive relationships between the extent of diversion and the analytical measures of inpatient capacity and ED size and use them to form hypotheses for our empirical model.

We formulate a sample 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 aggregated annual data on a cross-section of California hospitals. We use structural factors such as ownership, location, trauma center status to explain the decision to use ED and operational factors derived from the analytical models to explain the extent of diversion conditional on the decision to use it. In order to make a fair comparison, we also empirically test the basic specification (where the operational variables are raw measures of the number of beds in the ED and the inpatient department and the occupancy of the inpatient department) described above within the framework of the sample selection model.

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

1. Our use of queueing models to derive empirically testable hypotheses is novel. It provides a theoretical basis for deriving measures of inpatient capacity and ED size in a cross-sectional sample of hospitals with different hospital sizes. We find that the measures of inpatient capacity and ED size derived under heavy traffic approximation better describe the data than those derived under fluid approximation. We also find that coefficients for most variables in the basic (non-queueing) empirical model are not statistically significant thus further demonstrating the value of the analytical model to derive appropriate measures.
2. To our knowledge, this is the first cross-sectional study to investigate the drivers of ambulance diversion in hospitals. While previous single hospital studies have used longitudinal data to estimate the impact of inpatient occupancy and ED size on extent of diversion, such rich data is not available for a cross-section of hospitals. Consequently, we resort to using aggregate

annual data and employ analytical models to derive meaningful relationships at that aggregate level.

3. Ours is also the first study to highlight theoretically and then document empirically the interaction between the capacity 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 spare capacity of inpatient beds on the extent of diversion is decreasing in the size of the ED and vice versa. It is not possible to observe this interaction in single hospital studies due to lack of meaningful variation in ED size.
4. 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 not significantly associated with the utilization of the inpatient department.
5. Our empirical findings suggest that it is not possible to generalize the findings from single hospital studies to a larger population and that a “one size fits all” policy aimed at improving inpatient capacity might not reduce the extent of ambulance diversion across the board.

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 6. Section 8 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 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 an inpatient bed (The Lewin Group 2002).

### **3. Literature Review**

This paper contributes to the study of ED operations in operations management and emergency medicine literature from both theoretical and empirical perspectives. The queueing model spanning the emergency department (ED) and the inpatient department contributes to the theoretical

literature (Section 3.1), which has traditionally focused either on decisions outside the hospital system such as ambulance planning or those within the ED such as staffing and bed planning. Our cross-sectional empirical analysis contributes to the emergency medicine literature on identifying operational drivers of ambulance diversion, which has hitherto focused on single hospital studies and surveys (3.3). Theoretical analysis of this system comprising the ED and the inpatient department allows us to derive appropriate measures of inpatient capacity and ED size, which have been previously identified in the emergency medicine literature as the main drivers of ambulance diversion. Our empirical analysis also contributes to the scant literature on the empirical estimation of queueing systems in the operations management literature (3.2).

### 3.1. Theoretical models of ED operations

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. Vassilocopoulos (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.

Vassilopoulos (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.

### **3.2. Empirical estimation of queueing models**

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, formulation of the empirical model (specifically the definition of independent variables) is guided by the theoretical analysis of the underlying queueing network in different regimes. Indeed, our empirical results show that raw measures of inpatient utilization and ED size do not show meaningful association with extent of diversion. 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.

### 3.3. Empirical investigation of ambulance diversion

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 longitudinal 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 their data suggests that the time spent by hospitals on diversion is higher for larger hospitals. 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 approximations of the queueing model, which allow us to appropriately redefine the measures of ED size and inpatient occupancy by explicitly accounting for the size of the hospital.

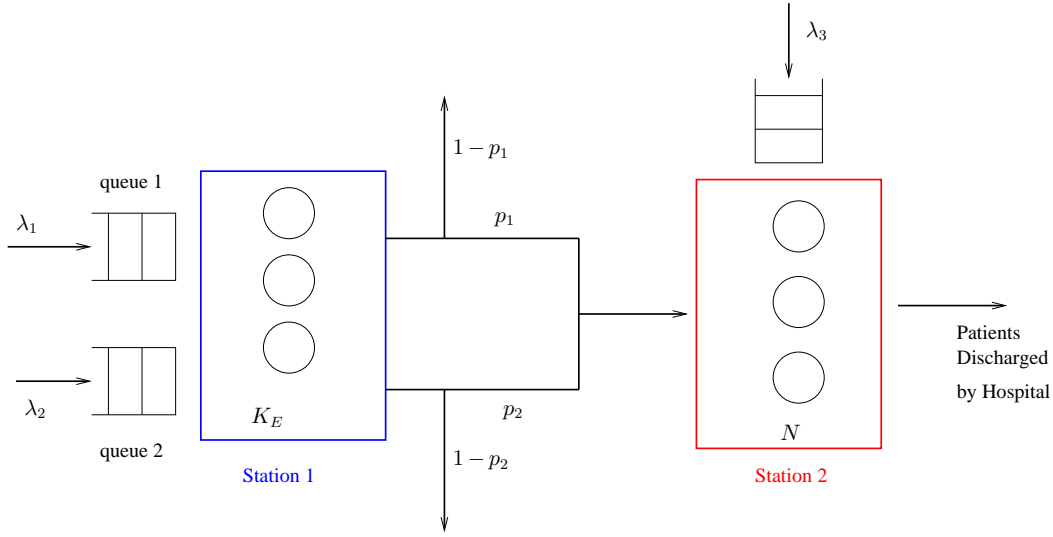
#### **4. Model Formulation**

Our primary objective of developing the analytical model in this paper is to derive appropriate measures of ED size and inpatient capacity and to form empirically testable hypotheses between these measures and the extent of ambulance diversion. Hence, given the aggregate nature of our data (described later) and foreseeable analytical complexity, we formulate a simplified steady-state stationary queueing model to capture the essential elements of the patient flow in the ED and the inpatient department. We develop two approximations for this model - heavy traffic and fluid - that are appropriate for large and busy systems and derive two set of empirically testable hypotheses.

We use the heavy traffic approximation to model unpredictable variation, i.e., stochastic variation in interarrival and service times while the fluid approximation is used to model predictable variation, i.e., variation in average arrival rates over the course of a day. Each approximation results in different relationships between the main operational parameters and the systems performance. Our objective in developing these models is not to accurately simulate the detailed dynamics of the ED patient flow but rather to provide explanation of the aggregate performance data.

##### **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.

**Figure 1** A Two-Station Queueing Model

**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  $\lambda_a$  and  $\lambda_w$ , 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  $N_2$  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  $\lambda_n$  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 continue 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  $m_2$ .

## 5. Analysis and approximations

We focus on two performance measures in our analysis: 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. 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 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.

### 5.1. Heavy Traffic Approximation

While we can estimate the delay probability and the fraction of time on diversion by simulation, it is not possible to provide their analytical characterization or to derive appropriate measures of inpatient capacity and ED size in a manner that allows us to construct testable hypotheses. In order to achieve this, we next approximate the queuing dynamics by their heavy traffic limits. As the name suggests, this approach involves approximating the original queuing 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 shown to be good approximations for the performance measures of the system of interest provided its traffic intensity is sufficiently close to 1.

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

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

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

(A3)  $\sqrt{n}(1 - \rho^n) \rightarrow \beta$  for some constant  $\beta$  as  $n \rightarrow \infty$ , where  $\rho^n = \lambda^n/(n\mu)$  is the traffic intensity of the  $n^{\text{th}}$  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 increases at a 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  $\mu$  and indexed by  $n$  such that:

(B1) The  $n^{\text{th}}$  system is an M/M/n/K<sup>n</sup> system with arrival rate  $\lambda^n$

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

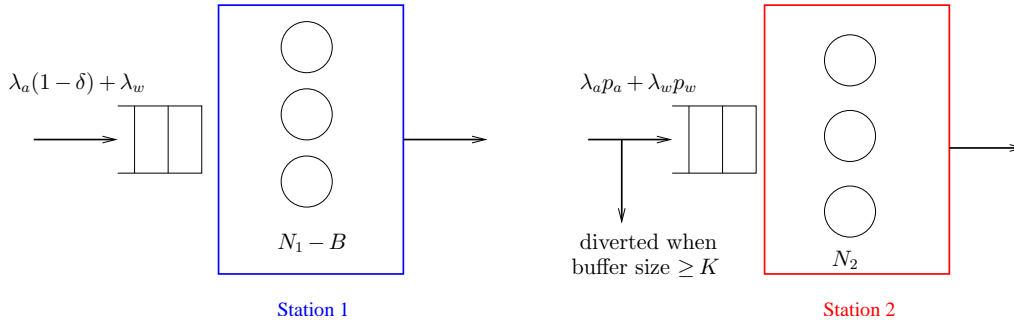
(B3)  $\sqrt{n}(1 - \rho^n) \rightarrow \beta$  and  $K^n/\sqrt{n} \rightarrow \kappa$  for some constants  $\beta$  and  $\kappa$  as  $n \rightarrow \infty$ .

**5.1.2. 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. Hence, we first introduce the following approximations to the queueing network to simplify the dynamics before deriving heavy traffic approximation:

- We approximate the overall arrival process to the ED by a Poisson process having rate  $\lambda_w + (1 - \delta)\lambda_a$ , where  $\delta$  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). We also validate the accuracy of this approximation when this condition is violated using numerical simulation (see Section 7

**Figure 2 A Simplified Queueing Model**

- 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.

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). Next, we apply the heavy traffic approximation to each of these single-station queueing systems.

**5.1.3. Application to the hospital model.** Using the basic approach described in section 5.1.1, we approximate the queue length processes at station 1 and station 2,  $Q_1$  and  $Q_2$  respectively, 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  $\delta$  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  $\delta$  by

$$\tilde{\delta}(N_2, \beta_2, \kappa) = \mathbb{P}(\tilde{Q}_2(\infty) \geq N_2 + K) = \frac{\beta_2 e^{-\kappa\beta_2}}{(\sqrt{N_2} - \beta_2) \left(1 - e^{-\kappa\beta_2} + \beta_2 \frac{\Phi(\beta_2)}{\phi(\beta_2)}\right)}, \quad (1)$$

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  $\beta_2$ .*
- (ii)  *$\tilde{\delta}(N_2, \beta_2, \kappa)$  is decreasing in  $\kappa$ .*

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  $\kappa$ ) will reduce the percentage of time on diversion  $\delta$ . However, as discussed earlier, increasing  $\kappa$  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  $\kappa$ ).

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

$$\tilde{P}_d = \mathbb{P}(\tilde{Q}_1(\infty) \geq N_1 - B) = \left[ 1 + \frac{\tilde{\beta}_1 \Phi(\tilde{\beta}_1)}{\phi(\tilde{\beta}_1)} \right]^{-1}, \quad (2)$$

where  $\tilde{\beta}_1$  depends on  $N_1, N_2, \beta_2$ , and  $\kappa$ . 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  $N_1, N_2, \beta_2$  and  $\kappa$ . The next Proposition formalizes the intuition that higher  $\kappa$  yields higher delay probability  $P_d$ .

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

Proposition 2 implies that for any  $N_1, N_2, \beta_2$  there is a unique  $\kappa$  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  $\frac{N_1}{\sqrt{N_2}}$ .

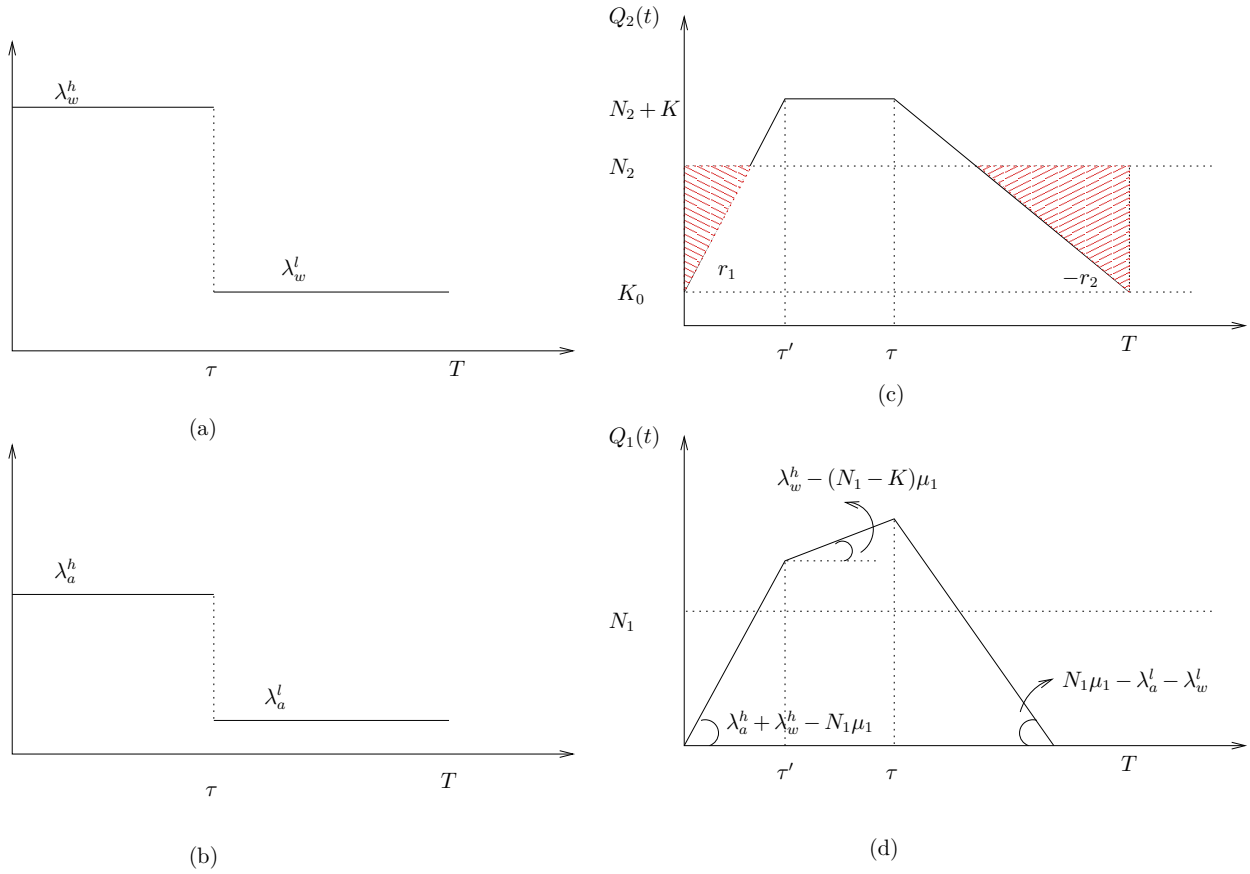
To summarize, the analysis of our queueing network model using heavy traffic approximation allows us to formulate the following hypotheses regarding how the fraction of time spent by the ED on diversion depends on the key structural characteristics of the system:

1. The fraction of time spent on diversion is decreasing in the spare capacity of the inpatient department, appropriately normalized for its size ( $\beta_2$ ), 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 (1).

## 5.2. Fluid Approximation

In this subsection, we develop a fluid approximation for the two-station queueing model described in Section 4 as an alternative to the heavy traffic approximation developed in Section 5.1. In this approximation, as the name suggests, we approximate the stochastic arrival (ambulance and walk-in) and service (ED and inpatient department) processes by deterministic and continuous flows. Simplifications arising from this approximation allow us to focus on an alternative explanation of diversion namely time varying arrival pattern for ambulances and walk-ins characterized by a peak

**Figure 3** Arrival rate pattern and inventory buildup for fluid model



period and an off-peak period, which results in ED experiencing intermittent periods of overloading and underloading.

We consider a period of length  $T$  (say 24 hours) that repeats itself. The peak arrival period is from  $t = 0$  to  $t = \tau$  and the off-peak period is from  $t = \tau$  to  $t = T$ . The arrival rates for ambulances and the walk-ins during the peak period are given by  $\lambda_a^h$  and  $\lambda_w^h$  respectively. Similarly, the arrival rates during the off-peak period are given by  $\lambda_a^l$  and  $\lambda_w^l$ . This is depicted in Figures 3(a) and 3(b). We approximate the service rate of the ED by  $N_1\mu_1$  when not on diversion and by  $(N_1 - K)\mu_1$  when on diversion, where  $K$  denotes the diversion threshold based on the blocked beds as before. We approximate the service rate in the inpatient department by  $N_2\mu_2$ . In order to ensure tractability and to eliminate cases which clearly do not correspond to the data, we make the following assumptions about the service and arrival rates:

- We assume that capacity in the inpatient department is sufficient to serve the arrivals during

the off-peak period but not enough to serve all the arrivals during the peak period, i.e.,  $\lambda_a^l p_a + \lambda_w^l p_w < N_2 \mu_2 < \lambda_a^h p_a + \lambda_w^h p_w$ . Note that if this condition is not true, the system would be always either underloaded (hence no diversion) or overloaded (hence always diversion), which are not observed in our data.

- We assume that capacity in the inpatient department is enough to serve all walk-in patients, i.e.,  $N_2 \mu_2 > \lambda_w^h p_w$ . Clearly, diversion is not useful as a mechanism to mitigate overcrowding in the absence of this condition.

- We assume that capacity in the ED is large enough to serve all those patients who would get admitted during the peak period, i.e.,  $N_1 \mu_1 > \lambda_a^h p_a + \lambda_w^h p_w > N_2 \mu_2$ .

**5.2.1. Diversion Probability.** Similar to our approach for the heavy traffic approximation, in order to derive an expression for the diversion probability, we first consider the dynamics in the second stage for a given diversion threshold  $K$ . Let  $Q_2(t)$  denote to be the number of patients that have completed service in the ED but are yet to complete service in the inpatient department including the boarding patients (Figure 3(c)).  $K_0$  ( $< N_2$ ) denotes the number of patients in the inpatient department at time 0 and let  $\tau'$  represent the first time that diversion starts, i.e.,  $\tau' = \min\{t : Q_2(t) = N_2 + K\}$ .

From  $t = 0$  to  $t = \tau'$ , i.e., when not on diversion, the nominal arrival rate to the inpatient department can be approximated by  $\lambda_a^h p_a + \lambda_w^h p_w$  since ambulance arrivals receive priority over walk-in arrivals and we assume that  $\lambda_a^h p_a \gg \lambda_w^h p_w$  and  $N_1 \mu_1 > \lambda_a^h p_a + \lambda_w^h p_w > N_2 \mu_2$ . Thus, from  $t = 0$  to  $t = \tau'$ , the number of patients in the inpatient department including the boarding patients increase at a rate of  $r_1 = \lambda_a^h p_a + \lambda_w^h p_w - N_2 \mu_2$ . Thus,

$$N_2 + K = K_0 + r_1 \tau'. \quad (3)$$

Next, consider the diversion period, i.e., from  $t = \tau'$  to  $t = \tau$ . Since the fluid model does not capture individual patient level dynamics, we model diversion as a fraction of ambulances that are not accepted by the ED, using a parameter  $0 < \delta' < 1$ . Thus, the nominal arrival rate to the

inpatient department during diversion is  $\lambda_a^h p_a (1 - \delta') + \lambda_w^h p_w$ . Since  $Q_2(t)$  is equal to  $N_2 + K$  at all times during the diversion, the flows in and out of the ED must be balanced. Hence,

$$\lambda_a^h p_a (1 - \delta') + \lambda_w^h p_w = N_2 \mu_2. \quad (4)$$

At time  $\tau$ , when the peak period ends and arrival rates jump from high to low, the number of patients  $Q_2(t)$  depletes at a rate of  $r_2 = N_2 \mu_2 - \lambda_a^l p_a - \lambda_w^l p_w$  and drops below  $N_2 + K$  thereby ending diversion. Since time  $T$  is also time 0 of next decision cycle,  $Q_2(T) = Q_2(0) = K_0$ . Note that balancing the flow between times  $t = \tau$  and  $t = T$ , we obtain

$$N_2 + K - K_0 = r_2(T - \tau). \quad (5)$$

Now, we can approximate the diversion probability as  $\delta = \delta'(\tau - \tau')/T$  since the ED diverts a fraction  $\delta'$  of the ambulances during the time between  $\tau'$  and  $\tau$ . Thus, combining (4), (3) and (5) we have

$$\delta = \frac{\lambda_a^h p_a + \lambda_w^h p_w - N_2 \mu_2}{\lambda_a^h p_a} \left( 1 - \frac{(N_2 + K - K_0)(r_1 + r_2)}{r_1 r_2 T} \right).$$

While the above is a closed-form expression for the diversion probability  $\delta$ , it contains several parameters that are not observable in the data. Hence, we use the fact that the utilization of the inpatient department is given by  $\rho_2 = 1 - \frac{\left( \frac{(N_2 - K_0)^2}{2r_1} + \frac{(N_2 - K_0)^2}{2r_2} \right)}{TN_2}$ . In addition, we restrict ourselves to analyzing the diversion probability  $\delta$  for a sequence of hospitals where the arrival rate scales linearly with the size of the inpatient department, i.e.,  $\lambda_i^j = \beta_i^j N_2, i \in \{a, w\}, j \in \{h, l\}$ .

Making these two substitutions, we obtain:

$$\delta = \frac{\beta_a^h p_a + \beta_w^h p_w - \mu_2}{\beta_a^h p_a} \left( 1 - \sqrt{\frac{2(1 - \rho_2)}{T}} - \frac{K}{N_2 T} \right) \quad (6)$$

It is clear from the above expression that  $\delta$  is decreasing in  $\theta_2 = \sqrt{1 - \rho_2}$  and decreasing in  $\kappa' = \frac{K}{N_2}$  where we use  $\kappa'$  to distinguish it from  $\kappa$  used in the previous section. Note that if the arrival rate increases less than linearly with the size of the hospital, the diversion probability will tend to zero for very large hospitals. On the other hand, if the arrival rate increases more than linearly with the size of the hospital, the diversion probability will converge to one for very large hospitals. Since we do not observe both of these scenarios in our data, we believe that choosing a linear scaling provides the best explanatory power to the model based on fluid approximation.

**5.2.2. Delay Probability.** Again, similar to the analysis of the heavy traffic approximation, we now consider the dynamics in the ED and derive the optimal threshold  $K$  such that the ED meets an exogenously specified delay probability.

Figure 3(d) depicts the build-up and draw-down of the number of patients that have not completed their service in the ED,  $Q_1(t)$ . We assume that the day begins with an empty ED, i.e.,  $Q_1(0) = 0$ . Before the diversion begins, i.e., from  $t = 0$  to  $t = \tau'$ , the number of patients in the ED builds up since  $\lambda_a^h + \lambda_w^h > N_1\mu_1$ . During diversion, i.e., from  $t = \tau'$  to  $t = \tau$ , rate of build up (or draw down) changes to  $\lambda_w^h - (N_1 - K)\mu_1$  due to simultaneous change in arrival rate (due to ambulance diversion) and service rate (due to blocking). After diversion, i.e., from  $t = \tau$  to  $t = T$ , the number of patients draws down at the rate of  $N_1\mu_1 - \lambda_t^a - \lambda_t^w$  until all patients are discharged. Note that in this model, the ED operates under two extreme regimes: arrivals to the ED either surely wait for a bed when  $Q_1(t) < N_1$  or surely do not wait for the bed when  $Q_1(t) > N_1$ . We define the fraction of time spent in the latter regime as the delay probability for this model.

In order to show that an optimal threshold  $K$  is increasing in  $N_1$ , we make two observations. First, for fixed  $N_1$ , an increase in  $K$  leads to an increase in the slope of  $Q_1(t)$  during periods of diversion, i.e., from  $t = \tau'$  to  $t = \tau$  thereby increasing  $Q_1(t) \forall t > \tau'$  and hence increasing the probability of delay. Second, for fixed  $K$ , an increase in  $N_1$  leads to a reduction in slope of  $Q_1(t)$  from  $t = 0$  to  $t = \tau$  and an increase in the slope for  $t > \tau$  thereby reducing  $Q_1(t) \forall t$  and hence decreasing the probability of delay. Combining the above two arguments, we can see that for a fixed delay probability, the threshold  $K$  is increasing in  $N_1$ .

In summary, analysis of the queuing network model using the fluid 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 the time spent on diversion is decreasing in the square root of the utilization of the inpatient department ( $\sqrt{1 - \rho_2}$ )
2. The fraction of the time spent on diversion is decreasing in the size of the ED, appropriately normalized for the size of the inpatient department ( $N_1/N_2$ )

Before turning to the empirical analysis, it is instructive to compare and contrast the two approximation approaches. Under both approximations, the extent of diversion is found to be decreasing in the relative size of the ED and decreasing in the normalized measure of the utilization of the inpatient department. However, the definition of these measures is quite different under both approaches. In the heavy traffic approximation, both of these measures are normalized using the square root of the number of inpatient beds. In the fluid approximation the utilization of the inpatient department is not normalized, whereas the size of the ED is normalized using the absolute number of inpatient beds.

The main reason for the difference in these relationships is the fact that the heavy traffic approximation accounts for the economies of scale effect, i.e., the impact of unpredictable variation reduces as the system size increases. This is not the case for the fluid approximation since it does not include unpredictable variation.

## 6. Empirical Analysis

Our analysis of the queueing model using the two approximations in section 5 provides us with empirically testable hypotheses regarding the relationships between the fraction of time spent on diversion and various measures of inpatient capacity and ED size. In this section, we attempt to understand which of these hypotheses are better supported empirically. We also estimate, for comparison, a basic empirical model using raw measures of inpatient capacity and ED size without any formal queueing theoretic foundations. Although these three empirical models differ in their independent variables, they share a common empirical modeling approach, which is described next.

### 6.1. 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 the 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:

$$y_{1i} = \alpha' \mathbf{Z}_i + \varepsilon_{1i} \quad (7a)$$

$$y_{2i} = \begin{cases} \gamma' \mathbf{X}_i + \varepsilon_{2i} & \text{if } y_{1i} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (7b)$$

(7a), 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, (7b), 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  $\varepsilon_1$  and  $\varepsilon_2$  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 (7a) is reformulated as follows by introducing another variable  $w_{1i}$  where  $w_{1i} = 1$  if  $y_{1i} > 0$  and  $w_{1i} = 0$  otherwise:

$$\mathbb{P}(w_{1i} = 1 | \mathbf{Z}_{1i}) = \Phi(\alpha' \mathbf{Z}_i) \quad (8)$$

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

Two methods are commonly used in the econometrics literature to estimate the selection model (7b) and (8). 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 (7b) 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 (8) 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 (7b), 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.

## 6.2. 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.

### 6.3. Measures

In Section 5, we derive hypotheses regarding the relationship between the outcome variable of interest, the fraction of time on diversion  $\delta$ , and various measures of inpatient capacity and ED size. For the heavy traffic approximation, we show that  $\delta$  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  $\frac{N_1}{\sqrt{N_2}}$ . Similarly, for the fluid approximation, we show that  $\delta$  is decreasing in the square root of the spare capacity of the inpatient department  $\sqrt{1 - \rho_2}$  and increasing in the absolute size of the ED  $N_1$ . Thus for empirical estimation, we need to first construct measures for the underlying variables  $\rho_2, N_2$  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  $\rho_2$  represents the utilization or occupancy of these beds and then use these to derive the measures for independent variables of interest.

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 ( $N_I$ )<sup>3</sup> and use it as a measure of  $N_2$ . Similarly, we calculate the theoretical utilization of staffed ICU beds  $\rho_I$  using the actual census of ICU patients reported and the staffed ICU beds  $N_I$  calculated above.

<sup>3</sup> staffed icu beds =  $\frac{\text{licensed ICU beds}}{\text{total licensed beds}} * \text{total staffed beds}$

We then use these measures to construct variables of our interest as follows. For the heavy traffic approximation, we define the relevant independent variables as  $REL\_ED\_SIZE \frac{N_1}{\sqrt{N_I}}$ , which reflects the size of the ED relative to the ICU and  $NORM\_ICU\_CAP (1 - \rho_I) \sqrt{N_I}$  which reflects the spare capacity in the ICU normalized for the size of the ICU. For the fluid approximation, we define the relevant independent variables as  $ED\_SIZE N_1$  and  $SQRT\_ICU\_CAP \sqrt{1 - \rho_I}$ . 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  $DIV\_HOURS>0$ ).

#### 6.4. Control variables

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).

### 6.5. 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 1 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 1 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

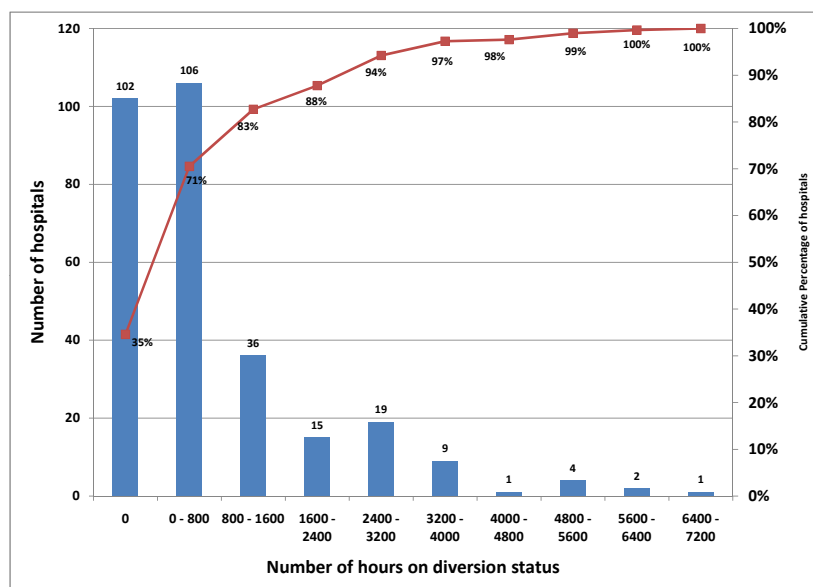
Note: Each cell contains the sample mean with the standard error in parentheses

Table 2 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 4. 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.

**Table 2** 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
$\rho_I$	0.66	0.31	0.00	2.05
REL_ED_SIZE	4.65	1.94	0.29	11.63
NORM_ICU_CAP	1.30	1.31	-2.99	6.65
RURAL	0.13			
INVESTOR	0.60			
GOVERNMENT	0.16			
TRAUMA	0.18			

**Figure 4** Distribution of ambulance diversion hours in study sample



## 6.6. Estimation and Results

We next test the hypotheses constructed based on the different queuing approximations. For comparison, we also test a basic model which uses different raw measures.

**6.6.1. Heavy traffic approximation.** We first test hypotheses which are based on the heavy traffic approximation of the queuing model. We use AMB\_DIV to denote the choice

of hospitals in (8) and DIV\_HOURS to denote the extent of diversion in (7b). In accordance with the results from section 5.1, we include variables NORM\_ICU\_CAP, REL\_ED\_SIZE and their interaction term NORM\_ICU\_CAP\*REL\_ED\_SIZE in (7b). 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 (8). Due to the presence of the interaction term REL\_ED\_SIZE\*NORM\_ICU\_CAP, we use centered variables REL\_ED\_SIZE\_CTR and NORM\_ICU\_CAP\_CTR 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 3. 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.

**Table 3** Estimation results for selection models based on heavy traffic approximation.

	MODEL I		MODEL II	
	LIML	FIML	LIML	FIML
	<b>LEVEL EQUATION</b>			
INTERCEPT	6.54*** (0.29)	7.33*** (0.16)	6.69*** (0.3)	7.35*** (0.15)
NORM_ICU_CAP_CTR	-0.06 (0.09)	0.01 (0.10)	-0.08 (0.08)	-0.04 (0.05)
REL_ED_SIZE_CTR	-0.07 (0.06)	-0.16** (0.04)	-0.01 (0.06)	-0.08*** (0.03)
(REL_ED_SIZE_CTR)*(NORM_ICU_CAP_CTR)	-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)	
	<b>CHOICE EQUATION</b>			
INTERCEPT	0.46*** (0.11)	0.55*** (0.08)	0.47*** (0.11)	0.47*** (0.09)
NORM_ICU_CAP_CTR	-0.04 (0.06)	-0.01 (0.04)		
REL_ED_SIZE_CTR	0.13*** (0.04)	0.08*** (0.03)		
(REL_ED_SIZE_CTR)*(NORM_ICU_CAP_CTR)	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. \*\*\* $p < 0.01$ ; \*\* $p < 0.05$ ; \* $p < 0.10$

For the “Level Equation”, the interaction term REL\_ED\_SIZE\*NORM\_ICU\_CAP is significant at 10% level indicating that REL\_ED\_SIZE and NORM\_ICU\_CAP 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 (1) which includes a product term  $\kappa\beta_2$ . The coefficients of NORM\_ICU\_CAP and REL\_ED\_SIZE are not significant

in Table 3. However, in the presence of interaction, these coefficients represent the conditional effect of these variables at the mean values 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 4. We find that coefficient of NORM\_ICU\_CAP is significant at 10% level for very high values of REL\_ED\_SIZE and coefficient of REL\_ED\_SIZE is significant at 10% for high values of NORM\_ICU\_CAP thus providing support for our theoretical results (Proposition 1 and 3.

**Table 4** Significance of simple slopes for REL\_ED\_SIZE and NORM\_ICU\_CAP in MODEL I.

	SLOPE OF NORM_ICU_CAP		SLOPE OF REL_ED_SIZE	
	LIML	FIML	LIML	FIML
REL_ED_SIZE_CTR = $-2*\sigma_R$	0.22 (0.19)	0.29 (0.22)		
REL_ED_SIZE_CTR = 0	-0.06 (0.09)	0.01 (0.10)		
REL_ED_SIZE_CTR = $2*\sigma_R$	-0.34* (0.20)	-0.27 (0.24)		
NORM_ICU_CAP_CTR = $-2*\sigma_N$			0.12 (0.12)	0.03 (0.14)
NORM_ICU_CAP_CTR = 0			-0.07 (0.06)	-0.16** (0.07)
NORM_ICU_CAP_CTR = $2*\sigma_N$			-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 REL\_ED\_SIZE): 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 NORM\_ICU\_CAP): 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 3 is very close to that in the LIML estimation but not significant. Examining the coefficients in the second and fourth column of Table 4, we again observe the phenomenon of shifting bottlenecks described earlier except that the coefficient of `NORM_ICU_CAP` was negative but not significant for large values of `REL_ED_SIZE`.

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 3. The results for both FIML and LIML estimation remain roughly similar except that the interaction term `REL_ED_SIZE*NORM_ICU_CAP` 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 `REL_ED_SIZE`, `NORM_ICU_CAP` and `REL_ED_SIZE*NORM_ICU_CAP` 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.

**6.6.2. Fluid approximation.** We next estimate a model which is based on the fluid approximation of the queueing system. Similar to the previous specification, we use `AMB_DIV` to denote the choice of hospitals in (8) and `DIV_HOURS` to denote the extent of diversion in (7b). In accordance with the results from section 5.2, we include operational variables `SQRT_ICU_CAP` and `ED_SIZE` in (7b) and variables `RURAL`, `GOVERNMENT`, `INVESTOR`, `TRAUMA` in (8). Also similar to the previous section, we estimate two model variants with and without the operational variables in (8) and use two estimation methods LIML and FIML for each model. The results are shown in the table below.

First, note that the Inverse Mills Ratio is significant at 10% for Model II indicating some evidence of sample selection. Observing the coefficients in the level equation, we observe that none

**Table 5 Estimation results for selection models based on fluid approximation.**

	MODEL I		MODEL II	
	LIML	FIML	LIML	FIML
	<b>LEVEL EQUATION</b>			
INTERCEPT	6.97*** (0.51)	7.53*** (0.52)	7.03*** (0.3)	7.73*** (0.31)
SQRT_ICU_CAP	-0.56 (0.46)	0.98** (0.50)	-0.81* (0.42)	0.04 (0.13)
ED_SIZE	-0.03 (0.06)	-0.11 (0.07)	-0.00 (0.06)	-0.06 (0.06)
INVERSE MILLS RATIO	-0.72 (0.56)		-0.94* (0.54)	
	<b>CHOICE EQUATION</b>			
INTERCEPT	0.33 (0.33)	0.45** (0.23)	0.48*** (0.11)	0.44*** (0.09)
SQRT_ICU_CAP	-0.78*** (0.30)	-0.49** (0.22)		
ED_SIZE	0.09** (0.04)	0.06* (0.03)		
RURAL	-1.19*** (0.27)	-0.48*** (0.03)	-1.36*** (0.26)	-0.40*** (0.06)
TRAUMA	0.21 (0.22)	0.04** (0.02)	0.26 (0.21)	0.31* (0.16)
INVESTOR	0.63*** (0.21)	-0.25*** (0.02)	0.45** (0.20)	-0.16*** (0.06)
GOVERNMENT	-0.49* (0.3)	-0.29*** (0.03)	-0.48* (0.29)	-0.27*** (0.03)
Log Likelihood				

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 the variables are consistently significant across both specifications and both estimation methods. The only variable that is intermittently significant SQRT\_ICU\_CAP actually changes sign between the two estimation methods. For the FIML estimation method in Model I, the coefficient of SQRT\_ICU\_CAP is (non-significant, yet) positive, which is clearly counter to underlining assumption underlining the fluid model. As a result, we conclude that the measures derived under the fluid approximation do not have face validity in explaining the data. As a result, we exclude this specification from further analysis.

**6.6.3. Basic model.** In this section, we define a basic model specification where the operational variables that enter the level equation (7b) are simply the raw measures of ED\_SIZE  $N_1$ , ICU\_CAP  $1 - \rho_I$  and ICU\_SIZE  $N_I$ . In order to level the playing-field between the different models we do not treat this model as a naive model, and thus we do not change the strategic variables such as ownership, location and trauma center designation that enter the selection equation. Also, we again consider two model variants depending on whether operational variables enter the selection equation or not and employ both estimation procedures.

Similar to the results for the specification based on fluid approximation, we find that coefficients for most variables are not statistically significant. Moreover, the coefficient for ICU\_SIZE is positive for Model II for both estimation methods suggesting that hospitals with larger ICUs have higher extent of diversion. This is counterintuitive to the principle of statistical economies of scale for systems with congestion, which states that larger systems have lower delay for same level of congestion. Thus, we conclude that this base specification does not have face validity in explaining the data. Hence, we exclude it from further analysis.

## 6.7. Network effect

Results from several recent studies suggest that ambulance diversion is a network phenomenon, i.e., diversion at one hospital affects the congestion and hence consequently the diversion at the neighboring hospitals (Sun et al. 2006). There is also some anecdotal evidence of “defensive diversion”, wherein a hospital goes on diversion immediately after its neighboring hospital goes on diversion in anticipation of the increased load and congestion (Vilkes et al. 2004). Thus, it can be hypothesized that in the presence of such a network effect, the extent of diversion at a hospital does not depend solely on its own structural characteristics but also on the configuration of the network of hospitals around it. We added a measure of network density in our level equation to account for this effect. We calculated distance between each pair of hospitals in our sample and then calculated the number of hospitals in a five mile radius around each hospital. We then created a categorical variable depending this number: no hospitals (NET\_NON), less than five hospitals (NET\_LOW), between

**Table 6** Estimation results for selection models based on base specification.

	MODEL I		MODEL II	
	LIML	FIML	LIML	FIML
	<b>LEVEL EQUATION</b>			
INTERCEPT	5.90*** (0.60)	7.42*** (0.50)	5.82*** (0.45)	7.19*** (0.17)
ICU_CAP	-0.37 (0.33)	0.12 (0.37)	-0.51 (0.31)	-0.10 (0.08)
ICU_SIZE	0.01* (0.01)	0.01 (0.01)	0.01* (0.01)	0.00 (0.01)
ED_SIZE	0.00 (0.01)	-0.02 (0.02)	0.01 (0.01)	0.00 (0.01)
INVERSE MILLS RATIO	-0.76 (0.56)		-1.11** (0.53)	
	<b>CHOICE EQUATION</b>			
INTERCEPT	0.47*** (0.11)	0.53*** (0.08)	-0.72*** (0.27)	-0.31 (0.24)
ICU_CAP	-0.55*** (0.22)	-0.39** (0.20)		
ICU_SIZE	0.01* (0.01)	0.00 (0.01)		
ED_SIZE	0.03*** (0.01)	0.03*** (0.01)		
RURAL	-0.91*** (0.28)	-0.63*** (0.23)	-1.37*** (0.26)	-0.45*** (0.04)
TRAUMA	0.09 (0.25)	-0.02 (0.17)	0.36* (0.22)	0.01 (0.06)
INVESTOR	0.50** (0.21)	-0.24*** (0.03)	0.81*** (0.22)	0.18 (0.19)
GOVERNMENT	-0.47* (0.29)	-0.29*** (0.03)	-0.43 (0.30)	-0.33* (0.19)
Log Likelihood				

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

five and ten hospitals (NET\_MED) and more than ten hospitals (NET\_HIGH). The results for this specification are shown below where the base category is chosen to be NET\_HIGH.

Note that the interaction effect between NORM\_ICU\_CAP and ED\_SIZE is still significant in most specifications in the presence of the network effect. The coefficients in the selection equation also remain largely unchanged indicating that the robustness of our results. Moreover, the sign of the coefficients for the network variables is as expected: hospitals in less dense networks have lower

**Table 7 Estimation results for selection models based on heavy traffic approximation with network effect.**

	MODEL I		MODEL II	
	LIML	FIML	LIML	FIML
	<b>LEVEL EQUATION</b>			
INTERCEPT	6.66*** (0.42)	7.63*** (0.32)	6.77*** (0.42)	7.59*** (0.18)
NORM_ICU_CAP_CTR	-0.05 (0.08)	0.02 (0.09)	-0.07 (0.08)	0.01 (0.06)
REL_ED_SIZE_CTR	0.07 (0.06)	-0.11 (0.08)	0.08 (0.06)	-0.05** (0.03)
(REL_ED_SIZE_CTR)*(NORM_ICU_CAP_CTR)	-0.12*** (0.04)	-0.04 (0.04)	-0.12*** (0.04)	-0.03* (0.02)
NET_NON	-1.19*** (0.47)	-0.72 (0.45)	-1.16*** (0.47)	-0.17 (0.23)
NET_LOW	-0.68* (0.39)	-0.59 (0.38)	-0.67* (0.39)	-0.25*** (0.07)
NET_MED	-0.08 (0.41)	-0.34 (0.34)	-0.08 (0.41)	-0.39*** (0.08)
NET_HIGH	-0.07 (0.06)	-0.16** (0.04)	-0.01 (0.06)	-0.08*** (0.03)
INVERSE MILLS RATIO	-0.22 (0.54)		-0.45 (0.55)	
	<b>CHOICE EQUATION</b>			
INTERCEPT	0.46*** (0.11)	0.52*** (0.10)	0.47*** (0.11)	0.47*** (0.09)
NORM_ICU_CAP_CTR	-0.06 (0.06)	-0.03 (0.05)		
REL_ED_SIZE_CTR	0.12*** (0.04)	0.10*** (0.04)		
(REL_ED_SIZE_CTR)*(NORM_ICU_CAP_CTR)	0.01 (0.03)	0.02 (0.02)		
RURAL	-1.27*** (0.27)	-0.77*** (0.31)	-1.36*** (0.26)	-0.39*** (0.05)
TRAUMA	0.26 (0.22)	0.22 (0.14)	0.26 (0.21)	0.14*** (0.05)
INVESTOR	0.63*** (0.21)	-0.03 (0.18)	0.45** (0.20)	-0.26*** (0.05)
GOVERNMENT	-0.46 (0.3)	-0.31* (0.19)	-0.48* (0.29)	-0.31*** (0.09)
Log Likelihood				

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

extent of diversion after controlling for operational variables such as inpatient capacity and ED size. While there has been extensive discussion about the network effect in the literature, there have been very few studies that have empirically documented the network effect (Sun et al. 2006, Vilkes et al. 2004) and none of the studies to our knowledge considers the joint effect of hospital characteristics and network characteristics on the extent of diversion.

## 6.8. Discussion and Limitations

Preceding analysis highlights the complementary value of analytical and empirical models. Specifically, our analysis of the two station queueing model provided the theoretical foundation for the measures of inpatient capacity and ED size that were used in the empirical models. On the other hand, empirical analysis provided a method of comparing the validity of the different analytical models. The empirical specifications obtained from different analytical models are not nested, which precludes us from making rigorous statements about model fit. However, we can conclude that the measures derived using heavy traffic approximation have better face validity as compared to those derived using fluid approximation and the raw measures themselves.

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. However, this creates a bias against finding significant effect on inpatient occupancy on ED overcrowding and ambulance diversion.

Our measure of staffed ICU beds was calculated based on the licensed ICU beds and the overall ratio of staffed and licensed beds in the hospital. We tested the robustness of our results using alternative measures of inpatient capacity. Using licensed ICU beds instead of staffed ICU beds did not dramatically change our results. However, using all the beds designated to general acute care instead of only intensive care resulted in loss of significance for most of our variables indicating that it is the ICU occupancy, which is the main driver for crowding in the ED.

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 allow us to assess the face validity of different measures of inpatient capacity and ED size derived using the different queueing approximations.

## 7. Simulation

In this section, we simulate the QED model described in section 4 to assess the accuracy of the simplifications and approximations made during our analysis in section 5. The performance measure

of primary interest is the fraction of time on diversion but we also assess the accuracy of the delay probability estimation. We compute the fraction of time on diversion using the two approximations 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  $K = 7$ , and  $N_1 = 15$ , the arrival rates are  $\lambda_a = 6$ ,  $\lambda_w = 12$ , and  $\lambda_n = 5$ , the service times are  $m_1 = 0.75$ , and  $m_2 = 2$ . We also assume that the likelihood that customers arriving to the emergency department continue to an inpatient department are  $p_a = 0.6$  and  $p_w = 0.005$ . Table 8 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

**Table 8** Estimation of fraction of time on diversion: approximation vs simulation

$N$	$\rho_2$	Approximation	Simulation	Difference
18	0.962	0.0805	0.0829	3.00%
20	0.866	0.0468	0.0486	3.76%
22	0.787	0.0269	0.0281	4.16%
24	0.722	0.0156	0.0163	4.39%
26	0.666	0.0091	0.0095	4.62%
28	0.619	0.0054	0.0058	7.42%
30	0.577	0.0032	0.0035	7.52%

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.

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  $N = 18$ . Table 9 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

**Table 9 Estimation delay probability: approximation vs simulation**

$N_1$	$\rho_1$	Approximation	Simulation	Difference
15	0.7995	0.3584	0.3695	3.01%
16	0.7421	0.2387	0.2531	5.66%
17	0.6924	0.1581	0.1723	8.26%
18	0.6489	0.1040	0.1166	10.78%
19	0.6106	0.0680	0.0783	13.11%
20	0.5765	0.0442	0.0522	15.27%
21	0.5460	0.0286	0.0346	17.41%

**Table 10 Estimation diversion probability, approximation vs simulation, as a function of the boarding-beds-threshold**

$K$	Delay probability	Diversion Approximation	Diversion Simulation	Difference
5	0.2520	0.1034	0.1093	5.36%
6	0.3011	0.0907	0.0951	4.63%
7	0.3584	0.0805	0.0829	3.00%
8	0.4250	0.0720	0.0743	3.14%
9	0.5023	0.0649	0.0665	2.36%
10	0.5916	0.0589	0.0599	1.63%
11	0.6944	0.0537	0.0542	0.87%

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.

We also conducted a numerical study to validate the accuracy of the approximation when varying the threshold level at which the hospital begin diverting ambulances. We fix the parameters described above as well as the ICU size at 18 beds and the ED size at 15 beds. We vary the threshold and compute the delay probability and the diversion probability. Table 10 reports these probabilities as well as the accuracy of the diversion probability approximation. Consistent with our explanation above, as the hospital increases the threshold, the diversion probability decreases, yet the probability of being delayed increases.

Lastly, we conducted a numerical study that aimed at testing the sensitivity of the approximation to violation of the condition  $\lambda_a(1 - \delta)p_a \gg \lambda_w p_w$ . We fixed the parameters discussed above as well

**Table 11** Estimation diversion probability, approximation vs simulation, as a function of the probability of

admitting a walk-in patient				
$p_w$	$\rho_2$	Approximation	Simulation	Difference
0.05	0.613	0.0041	0.0042	2.01%
0.1	0.653	0.0057	0.0059	3.41%
0.2	0.733	0.0112	0.0117	3.96%
0.3	0.813	0.0220	0.0231	4.95%
0.4	0.893	0.0404	0.0433	6.72%
0.5	0.973	0.0681	0.0744	8.49%

as the size of the hospital to 30 beds and computed the diversion probability for different values of  $p_w$ . As we increase the probability above 0.2 the condition is violated. Table 11 reports the comparison between the approximation and the simulation and shows that even when the condition is violated the approximation is still fairly accurate. The main conclusion from this table is that while the assumption is required for the purpose of analytical tractability, even if the condition is not satisfied, the accuracy of the model is quite high.

## 8. 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 using two approximations for large and busy systems. Our analysis shows that measures of ED size and inpatient occupancy change depending on the approximation chosen, which provides us with the opportunity to test our predictions empirically. We find that the measures developed using heavy traffic approximation have better face validity in explaining the data than those based on fluid approximation. We also find that

the empirical specification based on raw measures of ICU spare capacity and ED size, i.e., base specification, is not supported by data thus highlighting the value of our theoretical analysis in developing appropriate measures of capacity and deriving meaningful empirical results.

For the heavy traffic approximation, we find that the extent of ambulance diversion is decreasing in the spare capacity of the ICU and decreasing in the size of the ED, where both are appropriately normalized by the size of the ICU. 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 has relatively higher spare capacity.

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.

Our work can be extended in several ways; the most promising being modeling of the network effect, i.e., the interaction between diversion decisions in neighboring hospitals. Our empirical analysis provides preliminary and indirect evidence for the existence of this effect. However, further analytical work could provide insights into the mechanisms that lead to the network effect. For instance, is the correlation between the extent of diversion at neighboring hospitals driven primarily by the physical overflow of ambulance arrivals, i.e., response to actual load or by defensive behavior of hospitals, i.e., response to anticipation of load. Understanding the relative contribution of these mechanisms to the overall level of diversion observed can guide the policy discussion towards more subtle issues such as how to design diversion policies to obtain benefits of capacity pooling and away from sweeping measures such as repealing diversion altogether.

## 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  $\rho_1$  denote the traffic intensity of station 1 (the ED). It is clear that

$$\rho_1 = (\lambda_w + (1 - \delta)\lambda_a)m_1 / (N_1 - B). \quad (9)$$

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

$$\beta_1 = \sqrt{N_1 - B}(1 - \rho_1). \quad (10)$$

Therefore, to obtain an estimate of  $P_d$ , we need estimates of  $B$ , the average number of boarding patients, and  $\rho_1$ , the traffic intensity of the ED.

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

$$\tilde{B}(N_2, \beta_2, \kappa) = \sqrt{N_2} \frac{1 - e^{-\kappa\beta_2}(1 + \kappa)}{\beta_2(1 - e^{-\kappa\beta_2} + \beta \frac{\phi(\beta_2)}{\Phi(\beta_2)})} \quad (11)$$

Next, from (9), we can approximate  $\rho_1$  as

$$\tilde{\rho}_1(N_1, N_2, \beta_2, \kappa) = \frac{\lambda_1(1 - \tilde{\delta}(N_2, \beta_2, \kappa)) + \lambda_2}{N_1 - \tilde{B}(N_2, \beta_2, \kappa)} m_1, \quad (12)$$

where  $\tilde{\delta}$  and  $\tilde{B}$  are given in (1) and (11) respectively. We can then substitute (12) and (11) in (10) 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  $\beta_2$ . To show  $g_1$  is nondecreasing, we look at the first order derivative:

$$g_1'(\beta_2) = e^{\kappa\beta_2/2}[\kappa(\sqrt{N_2} - \beta_2)/2 - 1] \quad (13)$$

$$= e^{\kappa\beta_2/2}(K\rho_2/2 - 1) \geq 0. \quad (14)$$

For  $g_2$ , it is easy to see that  $e^{\kappa\beta_2/2}\Phi(\beta_2)/\phi(\beta_2)$  is increasing in  $\beta_2$ . 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  $\beta_2$ . 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 (1) 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  $\kappa$ .

*Proof of Proposition 2* We can rewrite equation (11) 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  $\kappa$  because  $\frac{\kappa e^{-\kappa\beta_2}}{1 - e^{-\kappa\beta_2}} = \kappa / (e^{\kappa\beta_2} - 1)$  is decreasing in  $\kappa$ . Since  $\tilde{\delta}$  is decreasing in  $\kappa$ ,  $\tilde{\beta}_1$  is decreasing in  $\kappa$ . Therefore,  $\tilde{P}_d$  is increasing in  $\kappa$ .

*Proof of Proposition 3* It is equivalent to show that  $\kappa^*$  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, L. F. McCaig, and R. Valverde. Analysis of ambulance transports and diversions among us emergency departments. *AEM*, 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 queuing 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. On the choice between sample selection and two-part models. *Journal of Econometrics*, 72(1-2):197–229, 1996.
- S. F. Leung and S. Yu. Collinearity and two-stage estimation of sample selection models: Problems, origins and remedies. *Computational Economics*, 15(3):173–199, 2000.
- 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. *ECON*, 52(3):737–760, 1984.
- B. C. Sun, S. A. Mohanty, R. Weiss, R. Tadeo, M. Hasbrouck, W. Koenig, C. Meyer, and S. Asch. Effects of hospital closures and hospital characteristics on emergency department ambulance diversion, los angeles county, 1998 to 2004. *Annals of Emergency Medicine*, 47(4):309–316, 2006.
- 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.
- G. M. Vilkes, E. M. Castillo, M. A. Metz, L. U. Ray, P. A. Murrin, R. Lev, and T. C. Chan. Community trial to decrease ambulance diversion hours: The san diego county patient destination trial. *Annals of Emergency Medicine*, 44(4):295–303, 2004.
- W. Whitt. A diffusion approximation for the G/GI/n/m queue. *Operations Research*, 52(6):922–941, 2004.