Phase I of these investigations involved formulation of a conceptual model that would permit data collection and analysis germane to the problem of ambulance diversion. As preparation for this study, a wide range of data was collected and analyzed from hospital emergency departments (ED) across the state of Massachusetts. The focus was to examine the root causes of ambulance diversion, with an emphasis on the factors that contribute to the diversion of non-emergency patients from the ED. The study aimed to identify and analyze the interrelationships of hospital services and patient care pathways in order to better understand the factors contributing to ambulance diversion.

An operations management perspective suggested straightforward input-throughput-output analysis. Hospital utilization data provided by the Division of Health Care Finance and Policy was therefore used to analyze the interrelationships of hospital services. In addition to simple supply/demand imbalances for emergency care, diversion and utilization patterns suggested bottlenecks and backlogs related to the competition of emergency and non-emergency patients for similar resources. The interrelationships of hospital services then became the focus of attention and patient care pathways were explored with administrators from the two study hospitals.

Two paradigms for the quantitative study of interrelationships among hospital departments were considered. The first involved an analytical approach wherein each relationship was identified, its stochastic character estimated, and appropriate mathematical models applied. The second involved an simulation approach, wherein stochastic relationships were embedded into computer software that translated real patient flow inputs into utilization patterns. Computer simulation was ultimately selected as the route of choice because of its scalability and adaptability.

Phase II

Data Collection/Analysis Effort:

The study was performed at two hospitals in Massachusetts: Hospital A, a large tertiary academic hospital, and Hospital B, a medium-sized acute care community hospital. The following data were collected:

- 42 days of information covering:
  - 6000+ admissions
  - 8000+ ED visits
  - 2000+ staffing/capacity data points
  - 300,000+ patient movement/status data points

In order to analyze the relationship between diversion status and other factors within the hospital environment, all measures were split into observations at one hour increments. The study period of 42 days, with one hour segments, meant that the required collection of patient flow data well beyond the usual capabilities of contemporary hospital information systems.

Point-biserial coefficients of correlation, with diversion status as the binary variable, were examined against a variety of factors. Comparisons when using full hours of diversion versus partial hours as the "true" condition did not reveal significant differences, so partial diversion hours were evaluated as the "true" binary throughout the analysis for the sake of consistency.

It is important to note that in the real world, the decisions to commence or cease diversion status are, by their nature, highly subjective. Because the purpose of the study was to examine the root causes of diversion, we did not approach the task from the standpoint of critiquing or attempting to influence this inherent operational subjectivity. As a result, any such analysis is itself subjective to a certain degree. Because both hospitals straddled EMS regional borders and diversion rules vary by region, each hospital's data was used for the sake of determining diversion status rather than using centralized EMS data. Diversion codes were considered equally rather than separately analyzing the factors related to each individual diversion type.

Patterns of diversion were also examined as averages across the hours of the day and the days of the week in order to ascertain relevant hour of the day and day of the week patterns. This data analysis was performed separately for each of the hospitals.
Hospital A:

Diversion Pattern "Hospital A - Diversion Minutes by Hour"

- There were a total of 22 episodes of diversion which started and ended within the study, with an average length of 814 minutes. There was one episode that began prior to the study and ended after the study began and so is not included in this calculation, nor in any calculations which involve the beginning of diversion episodes.

- The hourly diversion pattern shows diversion is highest in the evening hours, settles back down during the early morning hours, and then stays steady until the mid to late afternoon (see Fig. 1).

- The goal of the project was to determine the drivers which create this pattern.

The following 3 hypotheses were tested to determine the drivers of diversions:

1. ED arrival rate is too high, leading to diversion when the ED becomes full.
2. ED processing of patients is too slow, causing backups that lead to diversion.
3. ED arrival and processing rates are fine, but there are not enough beds in the hospital to accommodate the admissions.

There are seven sets of data (see Fig. 2), each representing a different view of arrivals into the ED. The "Arrivals_0" category only includes new arrivals from the current hour. Each subsequent category, from "Arrivals_1" to "Arrivals_6" includes one more hour's worth added to the arrivals from the previous hour. Because the average length of stay was 340 minutes, 6 hours is used as the maximum lag. Correlation coefficients from each of these cumulatives to Avg Diversion Minutes by hour are as follows:

Arrivals_0 = -0.073
Arrivals_1 = 0.001
Arrivals_2 = 0.078
Arrivals_3 = 0.165
Arrivals_4 = 0.259
Arrivals_5 = 0.359
Arrivals_6 = 0.460

There is also a possible corollary to hypothesis #1, that overall ED census is a driver of diversion. When counting the non-boarding census and comparing it to diversion status, however, the resulting point-biserial coefficient (r = -0.051) makes clear that this potential explanation should be rejected as well.
again points towards examining hospital capacity as the primary determinate of diversion.

Census/Admissions/Discharges: Hospital B

The overall relationship between inpatient census and ED boarders in Hospital B was similar to that of Hospital A. A comparison of weekday census and ED boarders between the two study periods on 19th December is presented because scheduled demand played a far smaller role than that observed in Hospital A.

During the study period, there were 1,158 weekday unscheduled admissions (average: 38.6/day) and 208 weekday scheduled admissions (average: 6.9/day). This suggests very little operational flexibility in controlling the variability or timing of scheduled arrivals. This likely reflects a fundamental difference between most community hospitals and larger academic centers.

Hospital B Conclusions:

The findings at Hospital B are consistent with and reinforce those at Hospital A. Specifically, there was no evidence that ED process times were temporally or mechanistically related to ED diversion while the relationship between ED arrival rate and diversion was weak. Instead, the data suggest that factors outside of the ED are more important.

Phase II Summary:

Detailed flow analysis in two very different types of hospitals yielded similar findings with respect to the root cause of emergency department crowding and ambulance diversion. Neither increased patient inflow nor increased process times could be strongly related to diversion status. Instead, diversion was seen as an outflow problem, with busy emergency departments crowding as patients await transfer to crowded inpatient services. This problem is exacerbated in hospitals with significant volumes of scheduled admissions, since these necessarily compete for the same resources. The collision of scheduled and unscheduled patient flows results in diversion patterns that are predictable and reproducible. Because the volume of scheduled admissions is controllable, better understanding of this phenomenon may suggest means of decreasing diversion. If the experience here may be generalized, we conclude that institutions with small (or uncontrollable) scheduled patient flows will require addition of resources on the inpatient side if diversion is to be substantially reduced.
At the same time, the form of the agents' staffing does not changed for the days with different scenario for arrivals.

Number of agents (Retail)

<table>
<thead>
<tr>
<th></th>
<th>100</th>
<th>150</th>
<th>200</th>
<th>250</th>
<th>300</th>
<th>350</th>
<th>400</th>
<th>450</th>
<th>500</th>
<th>550</th>
<th>600</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Averages over two weeks (11.01.09-25.01.09)

There are two type of the number of agents with tags HOT and actual. The first number was calculated from the HOT data in the following way:

Number of agents (Net) * % of occupancy = agents (HOT).

The actual number was calculated in the following way:

Number of answered calls * (average service time + 30 seconds) = agents (actual).

This number describes the number of servers working without idle time.

Our conclusion from this figure and the following in the analysis below that the number of agents (HOT) is much bigger than it was in reality.

The figure above shows inconsistencies in forms of the offered load and agents' staffing. All values presented in this figure are averages of the offered load in two different scenarios and the number of agents calculated in each 30 minutes interval.

The following 4 figures show the average number of arrivals, agents, waiting time, answered calls over two weeks period (11.01.09-25.01.09)
An example of the best-case scenario.

We choose Sunday 18.01.09 as a day with the best performance characteristics between weekdays 18-15.01.09. Even in this day we see problems in the service in morning and evening hours.
The figure above shows inconsistency in forms of the arrivals and agents' staffing. All values presented in this figure are averages of the number of arrivals coming by two different scenarios and the number of agents, calculated in each 30 minutes interval.

Average service time changes dramatically during the day. We see that in the evening average service time is much higher than in the middle. This means that the offered load in the evening is higher than in the middle, but the capacity is the same. It is interesting to note how variability of arrivals and agents influences this variability of service times.
Staffing Time-Varying Queues:

Two Common Approaches:

**SSA** – Simple Stationary Approximation ($SSA, \lambda = 0.2$)

Could result in time-varying (highly oscillating) performance (utilization/service), which is undesirable.

M/M/N, with $\lambda = \lambda(t)$ at each time $t$.

Time-varying staffing levels, based on steady-state.

**PSA** – Point-Wise Stationary Approximation

with $\lambda = \lambda(t)$, long-run average number of arrivals.

Constant staffing levels, based on steady-state $M/M/N$.

**SSA** – Simple Stationary Approximation.

Two Common Approaches:

Staffing Time-Varying Queues:
Exponential having mean 0.1 (6 min)

Assume: Service and abandonment rates are both exponential having mean 0.1 (6 min).

Example: "Real Call Center"

(Two-hump arrival functions are common)

(Adapted from Green L., Kolese P., Soares J. for benchmarking)
Congestion (Queue, Wait)

ED = 0.9 Minutes
QED = 0.5 Seconds
QD = 0.1 Negligible


time

Average Queue
Average Wait

$\infty$ $M/M/\infty$

Note: The same transition rates as

$\lambda (N+1) / \mu N$

$\mu N - 0.1$

$\lambda / \mu$

For any $N$, transition rates for $\{L_t, t \geq 0\}$:

- $L_t$: number-in-system at time $t$ (Birth & Death)
- $\lambda$: service rate equals abandonment rate
- $\mu$: service rate equals abandonment rate with $M/N$ system

Emergency: A: Moderate (im)patience

Congestion (Queue, Wait)
Time-Varying Arrivals

\[ \frac{S - \text{excess service}}{z} \]

R\(1\) - the offered load at time 1, namely:

Fact: \( L \sim \text{Poisson}(R) \)

\[ R^1 \cdot \gamma + 1 = N \]

Extension: \( W / M / N + 1 \)

Square-Root Staffing: Motivation

\[ \{N < bM\}d \]

Fact:

\[ \{N < bM\} \]

With:

\[ R \sim \text{Normal}(R, \mu) \]

For \( R \) not too small:

\[ \frac{R^\gamma + Z}{R - N} \approx \frac{\infty / W / W (R, \mu)}{N} \]

\[ \{N < bM\}d \]

\[ \{N < bM\}d \]

\[ \{N < bM\}d \]

Square-Root Staffing: Motivation
Performance Measures

\[ N' = R' + f' \cdot \frac{N'}{R'} \]

The following Square Root Sharing rule:

\[ Service Grade \text{ in interval } t, \text{ which arises from the } \]

\[ Service Utilization \text{ in interval } t, \text{ calculated as the } \]

\[ Tail probability \text{ in interval } t, \text{ calculated as the probability that queue size equals or exceeds some threshold } (e.g. 3 times average queue length) \]

\[ Average queue length \text{ in interval } t, \text{ taken constant over all replicates and over the time-interval. The average length is averaged over the time-interval. The average length in interval } t \text{ is calculated as: } \]

\[ \text{Time interval } t \text{ average waiting time in interval } t \text{ calculated by the interval, calculated by the average waiting time of all customers arriving during the interval. } \]

\[ Delay probability \text{ in interval } t, \text{ calculated by the deviation probability of customers who are not served immediately. } \]

\[ Time-Varying Arrivals \]
Two-hump arrival functions are typical. Service and abandonment rates are both exponential having mean 0.1 (6 min.).

Example: “Real Call Center.”
Real Call Center: Empirical waiting time, given positive wait

(1) $\alpha=0.1$ ($QD$)  (2) $\alpha=0.5$ ($QED$)  (3) $\alpha=0.9$ ($ED$)
QED Staffing ($\alpha=0.5$)

Erlang A: Theoretical vs. Empirical

P(Wait>0) vs. $\beta$ (N=R+1/R)
$$P\{W > 0\} = \gamma$$

$$\theta$$

$$\gamma (N = \gamma + \theta)$$

$${\gamma}(0.1)$$

$${\gamma}(0.5)$$

$${\gamma}(1)$$

$${\gamma}(2)$$

$${\gamma}(5)$$

$${\gamma}(10)$$

$${\gamma}(20)$$

$${\gamma}(50)$$

$${\gamma}(100)$$

$${\gamma}(x)$$ describes the asymptotic probability of delay as a function of $${\gamma}$$ when

Here, $${\gamma}$$ and $${\theta}$$ are the abandonment and service rate, respectively.
Iterative Algorithm

**Inputs**

- System primitives:
  - arrival function, service-time distribution,
  - patience distribution (when relevant);
- Target delay probability $\alpha$;
- Time horizon $[0,T]$.

**Outputs**

- Staffing function, aiming at a delay probability $\alpha$ is over $[0,T]$.

**Starting point**: The *infinite-server heuristics* by Jennings, M., Massey, Whitt (1996)

Algorithm (cont.)

**Notation**: $\forall t \in [0,T]$ (*practically* $t=0$, $\Delta$, $2\cdot\Delta$, ...)

- $N_i(t)$ — staffing level at time $t$, determined in iteration $i=1,2,...$
- $L_i(t)$ — number in the system at $t$, under staffing function $s_i(t)$.

**Algorithm**:

1. $i=0$; $N_0(t) \equiv \infty$ (delay probability $=0$)
2. Evaluate the distribution of $L_i(t)$, using simulation.
3. Determine $N_{i+1}(t)$ as follows:
   
   $N_{i+1}(t) = \arg \min \{ c : P\{L_i(t) \geq c\} < \alpha \}, \quad 0 \leq t \leq T.$
4. Check stopping condition:
   
   - if $\|N_{i+1}(\cdot) - N_i(\cdot)\|_\infty \leq 1$, then $N_{i+1}(\cdot)$ is our staffing level;
   - else $i := i+1$, and go back to (2).
5. **Last iteration.** The algorithm converges to a Staffing Function $N_\infty(\cdot)$ least for which

\[ P\{L_\infty(t) \geq N_\infty(t)\} \leq \alpha, \quad 0 \leq t \leq T. \]