Using Queueing Theory to Increase the Effectiveness of Emergency Department Provider Staffing

Linda V. Green, PhD, João Soares, PhD, James F. Giglio, MD, Robert A. Green, MD

Abstract

Objectives: Significant variation in emergency department (ED) patient arrival rates necessitates the adjustment of staffing patterns to optimize the timely care of patients. This study evaluated the effectiveness of a queueing model in identifying provider staffing patterns to reduce the fraction of patients who leave without being seen.

Methods: The authors collected detailed ED arrival data from an urban hospital and used a Lag SIPP queueing analysis to gain insights on how to change provider staffing to decrease the proportion of patients who leave without being seen. The authors then compared this proportion for the same 39-week period before and after the resulting changes.

Results: Despite an increase in arrival volume of 1,078 patients (6.3%), an average increase in provider hours of 12 hours per week (3.1%) resulted in 258 fewer patients who left without being seen. This represents a decrease in the proportion of patients who left without being seen by 22.9%. Restricting attention to a four-day subset of the week during which there was no increase in total provider hours, a reallocation of providers based on the queueing model resulted in 161 fewer patients who left without being seen (21.7%), despite an additional 548 patients (5.5%) arriving in the second half of the study.

Conclusions: Timely access to a provider is a critical dimension of ED quality performance. In an environment in which EDs are often understaffed, analyses of arrival patterns and the use of queueing models can be extremely useful in identifying the most effective allocation of staff.

Keywords: administration, quality indicators, leave without being seen, quality improvement, emergency medicine

Several national reports have documented a growing demand for care from emergency departments (EDs) and a simultaneous decrease in the number of operating EDs. The result has been increased crowding, prolonged waiting times to be treated by an emergency provider (i.e., physician or physician assistant), and high percentages of patients leaving EDs without being seen. A recent study found that in 2001, 7.7% of the 36.6 million adults in the United States who sought care in a hospital ED reported trouble in receiving emergency care, and that more than half of these cited long waiting times as a cause. In addition, although the general pattern of demand is similar across the week, individual days are likely to experience different overall volumes and slight differences in the exact timing of peaks and valleys.
Among the foremost challenges in determining ED provider schedules is trying to match staffing levels to accommodate these changing demand levels. This is a difficult task for several reasons. First, even in the case of constant demand levels over the day, statistical fluctuations in individual patient arrival times and the variability in the time needed by a provider to treat patients can create long delays even when overall average staff capacity is greater than average demand. Second, the magnitude of delays is a nonlinear function of the demand or staffing level and is thus impossible to predict without the use of a queueing model. In an environment with time-varying demands, delays are likely to be even greater, particularly if staffing is not carefully adjusted based on the actual fluctuation of the arrival rate over the day. Furthermore, the level of staffing in any given interval affects delays in other staffing intervals, and the interaction effects are not predictable without the use of a model. Finally, staff levels at any given time may be constrained by organizationally mandated shift lengths and by the preferences of individual providers.

The primary goal of this study was to demonstrate the benefit of using a queueing model to construct ED provider staffing schedules that result in a more effective allocation of provider hours over the day and over the week.

METHODS

Study Design
We conducted a controlled trial with a “before/after” design. This study was deemed exempt from informed consent by the institutional review board.

Study Setting and Population
The study site is an urban ED in the Inwood neighborhood of northern Manhattan and has an annual census of approximately 25,000 patients. The population is 61% Hispanic, 18% African American, and 17% white. Twenty-five percent of patients arrive via ambulance. The admission rate for patients seen by a provider is 23%. At the time that the study began, staffing levels and shift schedules were identical for all days of the week using 55 provider hours per day. The ED was staffed with attending emergency physicians and physician assistants. There were no residents rotating through this ED during the study.

Study Protocol
The study examined the response of one ED measure of performance, left without being seen (LWBS), to a provider staffing reallocation based on queueing theory. Two 39-week periods, one before the staffing changes (August 26, 2002, to May 25, 2003) and one after the staffing changes (September 1, 2003, to May 30, 2004), were studied. Matching weeks were chosen to better control for seasonal variation in both volume and disease states. The intervals are not aligned by exact date to control for number of total days and days of the week. These date intervals result in exactly 39 complete weeks for both the before and after time intervals. The two periods of study are not contiguous; they are separated by a 14-week inter-vening period during which the staffing changes had started but were not yet fully implemented.

Queueing Model Description. An M/M/s queueing model was used to estimate the number of providers needed during each staffing interval (further explanation is provided in the Appendix, available as a Data Supplement at http://www.aemj.org/cgi/content/full/j.aem.2005.07.034/DC1). This model assumes a single queue with an unlimited waiting room that feeds into identical servers (e.g., providers). Arrivals occur according to a time-homogeneous Poisson process with a constant rate, and the service duration (e.g., provider time associated with a patient) has an exponential distribution. (These two assumptions are often called Markovian, hence the use of the two “M’s” in the notation used for the model).

One advantage of using the M/M/s model is that given an arrival rate, an average service duration, and the number of servers, formulae for performance measures such as the probability of a positive delay or the mean delay can be easily obtained and implemented on a spreadsheet.

Software packages that contain these formulae are widely available. The delay is measured from the time of the demand for service (e.g., patient registered in the ED) to the time at which service begins (e.g., a provider is available to treat that patient). It is important to note that the model’s delay predictions pertain only to waiting times due to provider unavailability and do not include any other possible delays before seeing a provider such as registration and triage times, which would have to be estimated independently.

Because the M/M/s model assumes that the arrival rate does not change over the day, actual service systems that have time-varying demands typically use this type of model as part of a SIPP (stationary independent period by period) approach to determine how to vary staffing to meet changing demand. The SIPP approach begins by dividing the workday into staffing periods (e.g., one, two, four, or eight hours). A series of M/M/s models is then constructed, one for each staffing period. Each of these period-specific models is independently solved for the minimum number of servers needed to meet the service target in that period. The service target might be a desired maximum mean delay or probability of delay standard. However, recent research has shown that the SIPP approach is often unreliable and that a simple modification, called Lag SIPP, is often more effective in identifying staffing levels that achieve the desired performance standard.

This is because in many service systems with time-varying arrival rates, the time of peak congestion significantly lags the time of the peak in the arrival rate. While the standard SIPP approach ignores this phenomenon, the Lag SIPP method incorporates an estimation of this lag and thus does a better job of identifying staffing levels to limit delays.

In this study, we used the Lag SIPP methodology, which was programmed using C as part of a prior research project, to identify provider staffing levels to achieve a given delay standard. The delay standard we chose was that no more than 20% of patients wait more than an hour before being seen by a provider. The use of one hour is consistent with the time standards associated with emergent and urgent patient groups used in the National
Hospital Ambulatory Medical Care Survey. The 20% criterion reflects the approximate percentage of nonurgent arrivals at the study institution.

**Measures**

For the performance analysis phase, patient disposition, arrival mode, age, gender, and length of stay were extracted from the Eagle 2000 registration database (Siemens Inc., Malvern, PA). Percent LWBS was defined as the total patients who LWBS divided by the total number of registered patients during the specified time period. Because patients are triaged before registration, it is possible that some patients who LWBS were not captured in our data collection.

A critical measure of ED performance related to provider staffing and patient throughput is the time from triage to the time to being seen by a provider. This measure was not recorded during the time period of this study. Instead, we used the strongly related measure, the proportion of patients who LWBS, as the determinant of the efficacy of the staffing changes suggested by our study. Previous studies have established a strong link between long ED delays and patients who LWBS. In addition, the proportion of patients who LWBS is itself an important measure of ED performance and quality of care. Several studies have concluded that patients who LWBS are sick and do require emergency care. One study has shown that up to 11% of patients who leave without being seen are hospitalized within one week and that 46% of patients were judged to require immediate medical attention.

**Data Analysis**

Data were extracted from the hospital’s admission database (Eagle 2000) using SAS version 9.1.3 (SAS Institute Inc., Cary, NC). ED hourly arrival data during 2002 were grouped by day of week. These data were used to construct the arrival rates needed as input to the queueing model. The queueing model also requires an average provider service time per patient, which must include the times of all activities related to a patient. These activities include direct patient care, review of radiographs and laboratory test results, telephone calls, charting, and speaking with other providers or consultants. At the time of the study, provider service times were not recorded. The only reference we were able to find in the literature that included such data reports an average service time of 24 minutes based on a prospective time study. For the purposes of our study, we used an average service time of 30 minutes based on the existing literature but also on productivity data and observation from the study site.

We constructed a multivariate logistic regression model using LWBS (0 or 1) as the dependent variable. The main independent variable was an indicator variable designating the original or new staffing (0 or 1). Daily mean total ED length of stay values, daily mean total visit values, and mode of arrival (ambulance vs. ambulatory) were used in a logistic regression model to assess the relationship between the staffing change and the odds of leaving without being seen. There was no change in nursing or technician staffing during the study period, and thus these variables were not included in the model. The model was applied to obtain odds ratios with 95% confidence intervals before and after adjustment for these potentially confounding factors. The Hosmer–Lemeshow goodness-of-fit test was used to test the multivariate logistic regression model fit. Statistical analysis was conducted using SAS version 9.1.3 (SAS Institute Inc.).

In addition to the analysis of the complete data set, the following subgroups were examined separately: weekdays, weekends, and the four-day subset of Saturday, Sunday, Monday, and Tuesday. Four hours were moved from both Saturday and Sunday and four hours were added to each of the weekdays. Therefore, by examining a four-day subset consisting of two weekend days and two weekdays, we were able to analyze the effect of reallocation of hours, both between days and within each day, without the confounding effect of the additional provider hours added to the schedule.

**RESULTS**

Queueing Model Analysis, Recommendations, and Resulting Insights

Linear regressions for arrivals for the 2002 calendar year showed that the ratio of the mean to the variance of the number of ED patient arrivals each half-hour by day of week was consistently close to 1. This supports the assumption of time-varying Poisson arrivals used by the queueing model. An examination of the hourly average arrival rates by day of week revealed that while the daily pattern of peaks and valleys was quite consistent, the overall average volume varied from a low of 63 patients per day on Saturdays to a high of 72 per day on Mondays. While this degree of variation indicated that the current policy of identical staffing levels for all days of the week was likely suboptimal, it was deemed impractical to have a different provider schedule every day. Therefore, we decided to use queueing analyses to develop two schedules: weekday and weekend. Figure 1 shows the aggregated average hourly ED visit rates for weekdays and weekends. We used these average hourly arrival rates and the estimated average provider time per patient of 30 minutes as input to the Lag SIPP routine.
to estimate staffing levels, based on two-hour staffing intervals, to achieve a maximum probability of 20% that a patient would wait more than one hour to be seen by a provider during any staffing interval.

The modeling results (Figures 2 and 3) indicated that a total of 58 provider hours were needed on weekdays to achieve the desired service standard, which represented an increase of three hours over the existing staffing level of 55 provider hours. Model runs for the weekend indicated that the target performance standard could be achieved with a total of 53 provider hours. In both of these cases, the queueing analyses suggested that the existing staffing pattern over the course of the day needed to be changed (see Figures 4 and 5 for the original and new staffing patterns). Specifically, it indicated that some provider hours should be switched from the middle of the night to much earlier in the day. This suggested change was further supported by the realization that more patients are impacted by staffing levels during high arrival rate intervals than during low demand levels. Therefore, implementing adequate staffing levels during the late morning, afternoon, and evening hours would have a greater positive effect on ED delays and LWBS levels than doing so during the middle of the night. A more subtle change suggested by the model was that the increase in staffing level to handle the morning surge needed to occur earlier than in the original schedule. The insights gained from these analyses became the guiding principles in developing new provider schedules.

Figure 2. Queueing model output of proposed staffing and patient arrival rate (weekdays).

Figure 3. Queueing model output of proposed staffing and patient arrival rate (weekends).
Development of New Schedules
The entire weekly staffing schedule was deconstructed and rebuilt based on the results of the queueing analyses. The resulting staffing requirement of 58 hours on weekdays to achieve the performance standard of provider contact within one hour for 80% of patients translated into 20 more hours on weekdays relative to the 55 hours per weekday that was then available. Although there was a 3% staffing increase approved at the time of the implementation phase of this study, this additional 12 hours still fell short of the recommended levels. The queueing model, however, facilitated a more logical placement of providers throughout the week and the day, including the movement of eight provider hours from the weekend to the weekdays.

Weekdays. Figure 4 illustrates the weekday staffing levels both before and after the change was made. The second provider on the overnight shift (10 PM to 6 AM) was moved to a daytime shift (2 PM to 10 PM). In addition, the noon to 8 PM shift was moved to a 10 AM to 6 PM shift based on the model results indicating a need to increase staffing earlier in the day. To better handle the high afternoon and evening volumes, an additional four hours were added to the
**Weekends.** Figure 5 illustrates the weekend staffing levels both before and after the change was made. The second provider on overnight shifts (10 PM to 6 AM) was eliminated. The noon to 8 PM shift was extended to midnight. This resulted in a net removal of four hours of provider time on both Saturday and Sunday and resulted in 51 provider hours (daily) for weekdays, two less than suggested by the queueing model.

**Outcome Results**

A total of 35,536 patients arrived to seek care in the ED during the 78 weeks examined. There was an increase of 1,078 patient visits (6.3%) during the implementation phase compared with the baseline period. Demographic characteristics of the patients in each group are shown in Table 1.

Table 2 contains our results on LWBS events. Considering the entire week, there were 258 fewer LWBS events (8.3% to 6.4% of total visits), despite the significant increase in ED visits. Isolating the four-day subset of the week for which there was only rearrangement of provider schedules and no net increase in provider hours, there were 161 fewer LWBS events; the proportion of LWBS events declined from 9.2% to 7.2%. This improvement is particularly noteworthy given that the number of visits for this subset increased by 548 patients (5.5%) between the “before” and “after” time periods.

On the weekends, when net provider hours were decreased by four hours each day, there was an increase of 87 LWBS events (6.7% to 8.2% of total visits). Weekday performance, when net provider hours were augmented by four hours each day, improved significantly, with a net decrease of 337 LWBS events; this represents a decrease in the proportion of LWBS events from 8.9% to 5.8%.

Table 3 contains the multivariate logistic regression results for the three covariates used in the models, all of which were significant. A one-hour increase of the average daily ED length of stay was found to increase the odds of LWBS events by 41%, while an additional patient arrival per day resulted in a 2.4% increase in odds of LWBS events. Arrival by ambulance decreased the odds of LWBS events by about 50%. These results were similar in all the subgroups: weekdays, weekends, and the four-day subgroup of Saturday to Tuesday.

**DISCUSSION**

Analytic models such as queueing models can never capture all characteristics of an actual operational setting. However, as has been demonstrated over many years and in an extremely broad variety of settings, models can be invaluable in providing decision support that greatly improves performance, particularly in complex environments. This study supports the usefulness of queueing models in guiding ED provider scheduling decisions. This is particularly true in EDs where resources are tight relative to demand because, in such situations, even small changes in staffing can have a dramatic impact on delays. Our study also demonstrates the need to collect data on and examine arrival patterns and to adjust daily staffing levels to assure that schedules are appropriate for what might be significantly different levels and patterns of demand across the week.

This work also highlights the importance of setting delay standards to obtain meaningful estimates of how much capacity is needed. An analytic model, in combination with a carefully developed, clinically appropriate delay standard, can provide an objective evaluation of what additional resources are required to meet a given standard of quality care. In the ED, timely treatment is most essential for emergent and urgent patients. Therefore, ideally, the standard used would reflect the time urgency associated with these types of patients and the queueing model would be priority based, reflecting the actual dynamics of the triage system. This was not done in this initial study because the patient information system did not accurately identify the triage status of patients. Future work to identify the best way to schedule additional provider hours that will become available this coming year will use the improved patient information system to produce a more refined analysis to reflect the triage classification.

**LIMITATIONS**

As mentioned previously, we did not have access to all of the data that are required for a queueing model. In particular, we had no data on the time providers spend with patients and had to estimate this based on the existing literature, observation, and judgment. We performed several analyses varying our estimate of this time and found that the resulting staffing levels were quite sensitive to this parameter. This highlights the need for an information system that can accurately capture these data. It is also important to note that the queueing model assumes that the time a provider spends with a patient is continuous, while in most cases, patient care is delivered in discontinuous time intervals. Examples included waiting for test results or breaks in care during patient care.
observation periods. The impact of these interruptions on delays is not clear, and there are no queueing models that incorporate this type of feature. Furthermore, without well-defined rules on how these service interruptions occur, they cannot be modeled.

We also did not have the ability to collect data on delays that patients experienced in being seen by a provider. Therefore, we could not directly validate the estimates produced by the queueing model. A new information system, implemented subsequent to this study, will enable the collection of these data in the future. In addition, due to constraints on the timing of provider shifts and personal preferences, the staffing schedules that were implemented were somewhat different from those that most closely aligned with the model’s suggestions. It is possible that our results would have been different had these constraints not existed.

Finally, we note that while the approach we used can be generalized to other hospitals, large EDs that have a more complex structure with regard to types of providers (e.g., attending physicians, residents, nurse practitioners) and locations and/or types of care (e.g., multiple districts and fast track areas) may require a more complex analysis. However, if patients and providers can be divided

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Full 7-day Week</th>
<th>Saturday to Tuesday</th>
<th>Weekdays</th>
<th>Weekends</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arrivals</td>
<td>0.0241 (1.02, 1.03)</td>
<td>0.0235 (1.018, 1.029)</td>
<td>0.028 (1.023, 1.034)</td>
<td>0.024 (1.015, 1.034)</td>
</tr>
<tr>
<td>Length of stay</td>
<td>0.345 (1.35, 1.48)</td>
<td>0.33 (1.31, 1.47)</td>
<td>0.31 (1.29, 1.44)</td>
<td>0.36 (1.31, 1.55)</td>
</tr>
<tr>
<td>Arrived by ambulance</td>
<td>−0.708 (0.44, 0.55)</td>
<td>−0.591 (0.48, 0.63)</td>
<td>−0.753 (0.41, 0.54)</td>
<td>−0.58 (0.56, 0.69)</td>
</tr>
</tbody>
</table>

For the noncategorical variables, arrivals and length of stay, the parameter estimates illustrated indicate the increase in the log odds ratio of leaving without being seen for each unit increase in the parameter. Length of stay is measured in hours (average daily length of stay), and arrivals refer to average daily patient arrivals. Arrival by ambulance is a binary categorical variable.
into independently operating segments, then each can be analyzed independently using the approach described in this report.

CONCLUSIONS

We believe that the queueing model, by providing a more rigorous and scientific basis for predicting patient delays in being seen by a provider, identified staffing schedules that reduced these delays and hence reduced the fraction of LWBS events. ED delays have been widely documented, and many hospitals have been engaged in efforts to reduce ED congestion, provide more timely care to patients, and reduce the percentage of LWBS events. Given the financial constraints that exist in most hospitals, it is increasingly important to find ways to improve performance with existing resources. This study illustrates how data analysis and queueing models can be used to identify staffing changes that can decrease the delays in being seen by a provider and, thus, the fraction of patients who leave without being seen, without necessarily increasing capacity. It also highlights the need to establish patient delay standards, preferably by triage class, and to establish information systems to collect and track data on provider service times and patient delays in seeing a provider.

The authors thank William T. Friedewald, MD, and Shing M. Lee for their consultation and advice relating to the statistical analysis.

References