What is queuing theory?
Queuing theory originated more than a century ago from the study of telephone delays and congestion. A simple queue is defined by a stream of arriving customers or tasks that are handled by a server. The goal of queuing models is to eliminate the disparity between the demand for service and the capacity to meet that demand. The concept is used in many service industries to strategize how to improve efficiency. For instance, a grocery store that struggles with long lines for check-out could add an express lane for customers with smaller amounts of items to reduce waiting time overall. In addition, call centers for customer service have used technology to improve caller wait times with the ‘virtual queue,’ which keeps the caller in line, but calls the customer back when the agent is free.

How does it work in the ED?
In the emergency department (ED), decreasing patient wait times is vital. Increased wait times lead to delayed diagnosis, poor patient satisfaction, and increased morbidity and mortality. The ED can be simplified to a ‘multiple server, single phase’ queue, in which patients wait in one line (the waiting room) for servers, including the triage nurse, the bedside nurse, and the physician.

In health care, queuing calculations are generally based on patient arrival rate, service rate (time for exam, tests, treatment), and the number of servers (number of providers and ancillary staff). Most models use the Poisson arrival process, which assumes patients arrive according to a random process. Arrival and service times can vary based on the time of day, the season, etc.

In the ED queue, interventions should focus on either reducing the server utilization or reducing variation. To reduce server utilization, the rate of service can be increased, with the goal of identifying wasteful elements and reducing or eliminating them. To reduce variation in service, the alignment of the staffing should first match demand before adding additional servers to the system. By predicting the average distribution of patient arrivals by hour, the staffing model can be adjusted to have more servers during high demand times. To reduce variation, the variation in arrival and/or the variation in service should be decreased. Reducing variation in arrivals can be difficult, as not much can be done to impact the timing of emergencies. Variation in service is usually related to issues of process, layout, supplies, equipment, and supporting services. For instance, a laceration repair might take longer if supplies are not readily available and stocked in the room where the procedure is being done. Even small adjustments to these issues can reduce service variation.

Can it work?
Queuing theory can be used to predict the effect of patient arrivals, treatment time, and ED boarding on the patients who leave without being seen (LWBS). One institution used a queuing model to analyze the ED flow model currently in place and found that a queuing model was able to predict the variation in patients who LWBS.

Lehigh Valley Health Network in Pennsylvania took the theory one step further. The institution used a queuing model to identify that the ED was understaffed during peak hours and overstaffed during non-peak hours. After aligning resource capacity with hourly demand, the hospital saw a reduction of length of stay by 20% and reduced walk-out rates by 58%.

Figure 1: Opportunities to Reduce Variation

Figure 2: How to Implement Queuing in the Emergency Department
- Determine patient arrivals/hour
- Understand the server staffing model
- Match patient demand with staffing
- Reduce service variation
- Analyze patient wait times/LWBS rates

ED’s are complex, especially in large, academic teaching hospitals. Relying solely on averages to determine forecasts can certainly affect the model. Accounting for residents and medical students in the formula can also impact predicted outcomes. However, queuing models can and should be considered in ED’s in order to improve efficiency and decrease wait times.
OPERATIONS MANAGEMENT

References: