Emergency Department Simulation Model

IEOR 174 Final Project

Ashlee Liu Sukhman Sidhu Shishira Somashekar Viplove Rahate

December 20, 2025

Contents

1 Background and Modeling Overview

Additional project materials are included with this report for completeness. A recorded presentation is available at:

https://drive.google.com/file/d/18GYUFD1MzviifT51JXk8jOEOpM44s31z/view?usp=sharing

The final presentation slides are available at:

https://docs.google.com/presentation/d/1oAPPZPurZwGdrn2PrIW7AxJPzEIydOb1559-5M7oy68/edit?usp=sharing

The complete simulation code, implemented in Google Colab, can be accessed at:

https://colab.research.google.com/drive/1IpEORgkqr4onbu1LRyl4e3lZe16yQYCe?usp=sharing

For convenience, the full code is also included in the Appendix of this report.

1.1 Problem Overview

Emergency departments have worked for decades to refine their processes, but methodologies are often disputed for there lack of efficiency. Many patients are burdened with long triage times, lack of resources, and unsuccessful visits. According to the US Department of Health and Human Services, "19.2 percent of persons had an emergency department wait time exceeding that recommended for any category in 2016. In addition, an article published in the National Institute of Health explains, "Although clinical staff, managers, and researchers are working to ease emergency care overcrowding, the problem continues to worsen. Emergency care overcrowding can weaken a hospital's adaptability to changes, which increases the likelihood of medical errors and adverse events and results in delayed treatments and longer waiting times." From this, it is clear that operational inefficiencies in hospitals need to be addressed quickly and systematically by looking at all levels of the larger system. This includes improper resource allocation, inadequate staffing, and inconsistency in treatment times.

These inefficiencies highlight a clear need for a simulation-oriented approach to identify bottlenecks. Our goal is to develop a simulation that can identify optimal staffing levels and provide actionable recommendations for hospital administrators. Through stochastic simulation, we seek to capture the randomness inherent in arrivals and other durations. These recommendations should prioritize reducing patient wait times and improving overall quality of care.

1.2 Project Objective

In this project, we will focus on modeling a mid-size emergency department. We will focus on how staffing levels influence resource utilization, bed wait times, and doctor wait times. The ultimate goal is to generate staffing recommendations that reduce patient waiting times and improve overall emergency department efficiency.

1.3 Emergency Department Simplified Process and Key Assumptions

The simulated emergency department follows a two stage process: (1) bed assignment and (2) doctor treatment. Patients arrive randomly, are assigned a triage level based on researched triage level frequencies, and must wait for an available bed before entering the queue for doctor treatment. To ensure the assumed triage level frequencies, wait times, and service times mimic a real world mid sized hospital we sourced our values from research published by the NIH and CDC.

The Center for Disease Control and Prevention Morbidity and Mortality Weekly Report (MMWR) published a graph that aggregates median wait and treatment times by triage level in 2010 and 2011. It is important to notes that about 17% of records were excluded from this analysis due to a wide range of reasons. All assumptions made on treatment and wait times were estimated from reading the grpah below. The graph is shown in Figure 1 at the end of this section.

Our facility capacity assumptions were based on a study of three suburban hospitals in the northeastern United States. The table we used had characteristics like yearly visits, physicians working during the study period, mean patients per shift for each site that was evaluated. The assumption we made from this cart was a facility capacity of 20 bers and 4 arrivals per hour. The chart is shown in Figure 2 at the end of this section.

The following research-driven distributions serve as the foundation for the simulation environment:

- **Triage Distribution:** Patient cases follow an empirical frequency distribution: Level 1 (0.41%), Level 2 (6.10%), Level 3 (40.20%), Level 4 (42.60%), and Level 5 (10.60%).
- **Wait Time Variability:** Each triage level is assigned distinct median wait times in minutes: Level 1 (5), Level 2 (17), Level 3 (25), Level 4 (30), and Level 5 (28)
- **Treatment Time Variability:** Each triage level is assigned distinct median treatment times in minutes: Level 1 (165), Level 2 (175), Level 3 (145), Level 4 (60), and Level 5 (45)

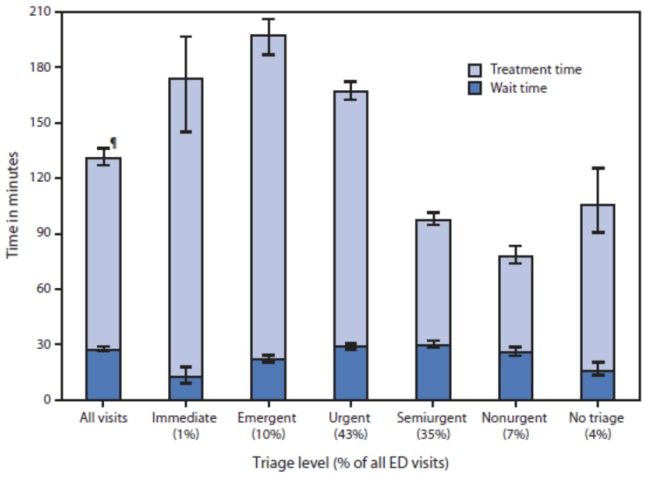

Figure 1: Center for Disease Control and Prevention Morbidity and Mortality Weekly Report (MMWR)

Table 1.

Characteristics of the study participants and sites evaluated

Characteristics	Site 1	Site 2	Site 3
Yearly visits	46 000	32 000	28 000
Shifts per day	6–7	4	3–4
Shifts evaluated	4444	2920	2458
Physicians working during the study period	22	17	25
Maximum number of concurrent physicians	3	2	2
Hours of single provider coverage	6	12	12
Hours of mid-level provider coverage	16	18	14
Mean patients per shift	15.0 (SD 4.7)	20.9 (SD 6.4)	13.2 (SD 3.8)
Percentage Emergency Severity Index 3 or more acute	73.1	76.2	76.9

Figure 2: Characteristics of 3 Suburban US Hospital (NIH)

While these factors remain constant across all simulations, the logic governing queue priority and physician staffing varies by model. These logical variations are explained in the following sections for the **(1) Baseline FIFO Model**, the **(2) Triage Priority Monte Carlo Simulation**, and the **(3) Doctor Staffing Capacity Model**.

2 Formulation, Variables, and Parameters

Building upon the foundation mentioned in Section 1.3, this section details the mathematical framework and stochastic processes that are the same in three models.

2.1 Arrival, Triage, and Service Time Modeling

Patient arrivals are modeled as a Poisson process with rate $\lambda = 4$ patients per hour ($\lambda = \frac{1}{15}$ patients per minute). Inter-arrival times are exponentially distributed with mean $1/\lambda$.

$$A_i \sim \text{Exponential}(\lambda).$$

Upon arrival, each patient is assigned a triage level $T_i \in \{1, 2, 3, 4, 5\}$ based on a fixed probability distribution p_k:

$$\mathbb{P}(T_i = k) = p_k, \quad \sum_{k=1}^{5} p_k = 1.$$

The values for p_k are derived from CDC/NIH frequency data.

Treatment times are dependent on the assigned triage level. Conditional on $T_i = k$, the service time S_i follows an exponential distribution with mean s_k:

$$S_i \mid (T_i = k) \sim \text{Exponential}\left(\frac{1}{s_k}\right).$$

As per our research-based assumptions, s_k is calculated by taking 15% of the total assigned treatment times, plus a mandatory 5-minute buffer to account for transition. This structure captures randomness in arrivals, variability in patient severity, and the workload variability of the doctor.

2.2 Resource Constraints

The system is limited by fixed physical and human resources, denoted by B (beds) and D (doctors). Consistent with a mid-sized facility, the system contains $B = 20$ beds and a baseline of $D = 2$ doctors.

Let $x_b(t)$ and $x_d(t)$ denote the number of beds and doctors in use at time t. At all times, system capacity constraints are enforced:

$$x_b(t) \leq B, \qquad x_d(t) \leq D.$$

These resource constraints, specifically number of doctors, are the primary influences of emergency department congestion, lengthy bed wait times, and doctor wait times.

3 Model Design

3.1 Model 1: Baseline FIFO Discrete-Event Simulation

3.1.1 Description

The first model evaluates emergency department performance using a based on an $M/M/k$ queuing system. The model follows a First-In, First-Out (FIFO) methodology, which serves as the primary operational constraint. The purpose of this model was to refine queuing methodology before incorporating triage priority. In addition, we were looking to understand the influence of triage priority on overall wait times.

The following constraints are applied to the model:

- **Priority Constraint:** In this specific model, the triage level is not taken into consideration for the ordering of who gets a bed or sees a doctor.
- **Sequential Processing:** Patients are processed in the order of their arrival, severity of their condition is ignored.
- **Stage 1 (Beds):** The system utilizes 20 beds ($k = 20$).
- **Stage 2 (Doctors):** The system utilizes 2 doctors ($k = 2$).

3.1.2 Initial Variables and Assumptions

The simulation parameters are defined by researched assumptions to mirror a 24-hour ED cycle:

- **Simulation Time:** 1,440 minutes (24 hours).
- **Arrival Rate (λ):** 4 patients per hour, modeled as a Poisson process.
- **Total Treatment Duration:** Total time in a bed is calculated by summing the "other service" time (nurse/rest), the "doctor treatment" (15% of total treatment time) time, and a mandatory 5-minute buffer for transition.
- **Distribution:** Inter-arrival and service times are sampled from an exponential distribution using median values specific to each triage level.

The probability density function (PDF) for the exponential distribution used is defined as:

$$f(t; \lambda) = \lambda e^{-\lambda t} \text{ for } t \geq 0$$

The total service time in minutes per patient is calculated as:

$$T_{service} = T_{otherservice} + T_{doctor} + 5$$

Metric	Result
Average Total Time in Bed	112.27 minutes
Average Bed Wait	0.00 minutes
Average Total Length of Stay (LOS)	112.27 minutes
Bed Utilization Rate	78.24%
Doctor Utilization Rate	92.79%

Table 1: Baseline Performance Metrics for FIFO Model

3.1.3 Simulation Results (Output Data, Shown Above)

The baseline performance of the FIFO model produced the following metrics based on the 24-hour simulation cycle:

3.1.4 Performance Analysis

The 24 hour simulation shows long wait times and high utilization rates, signifying a clear gap between throughput and emergency department readiness. The very low bed wait time is misleading because it is likely attributed to the FIFO logic. This logic is very unrealistic compared to how emergency departments actually work. In a real-world emergency department, staff would never prioritize a patient with a minor injury over a person with a life threatening condition just because of a small difference in arrival time. This model functions well as an operational baseline, but it lacks the ethical logic that is intended by assigning triage levels.

The high doctor utilization suggests that hospital is constantly performing a maximum capacity. This indicates that staff are likely overworked/unhappy, causing a degrading patient care experience. In addition, if there were to be a surge of arrivals due to unforeseen circumstances, the emergency department would struggle to operate.

3.2 Model 2: Triage Priority Monte Carlo Simulation

3.2.1 Description

Model 2 builds on Model 1 by incorporating triage priorities. The simulation tracks key patient-level metrics, including bed_time (time at which a patient is assigned a bed), bed_wait (time spent waiting for a bed), doc_wait (time waiting for a doctor after bed assignment), doc_start (time at which doctor service begins), and doc_end (time at which doctor service is completed). We run the simulation 50 times to capture the randomness and variability that is synonymous with emergency departments.

3.2.2 Simulation Results (Output Data)

Metric	Result
Average Bed Wait	0.00 minutes
Average Doctor Wait	5.67 minutes
Average Total Length of Stay (LOS)	95.14 minutes
Doctor Utilization Rate	44.27%
Bed Utilization Rate	27.37%

Table 2: Performance Metrics for Monte Carlo Simulation

3.2.3 Performance Analysis

The average bed wait time of 0.0 minutes and average bed utilization of 27.37% (roughly 5–6 of the 20 beds) show that beds are consistently available and do not constrain patient flow.

In contrast, patients experience an average doctor wait time of 5.67 minutes. Although doctor utilization is only 44.27%, the presence of waiting highlights the impact of stochastic arrivals and variable service times. From this, we can see that doctor availability is more likely the cause of delay than the number of beds.

3.2.4 Section Conclusion

The results from this 50 run Monte Carlo simulation model show that while bed capacity is sufficient, there is not an adequate amount of doctors. The Monte Carlo simulation was able to recognize that patterns in doctor wait times and

bed wait times were stable across the days. Overall, Model 2 provides an effective method for hospitals to understand bottlenecks in triage priority based queuing.

3.3 Model 3: Time-Varying Capacity Discrete-Event Simulation

3.3.1 Description

In Model 3, we will conduct a sensitivity analysis for five doctors split between two 12 hour shifts beginning at 9 AM and 9 PM respectively. The following combinations will be evaluated for number of AM doctors and PM doctors: (1,4), (2,3), (3,2), (4,1). For each combination, we use the same modeling approach as Model 2 and aggregates results for mean bed wait time, doctor wait time, and length of stay. Lastly, we plotted distribution of length of stay and doctor wait times across arrival times to understand model behavior.

3.3.2 Simulation Results (Output Data)

Day and Night doctors (day_docs)	Avg bed wait (minutes)	Avg doctor wait (minutes)	Avg total LOS (minutes)
(1, 4)	0.00	84.26	140.89
(2, 3)	0.00	4.05	96.29
(3, 2)	0.00	1.73	95.31
(4, 1)	0.00	5.75	96.54

Table 3: Performance Metrics by Staffing Configuration

3.3.3 Performance Analysis

For all four staffing combinations, the bed wait time is 0. Having only one doctor in either of the shifts resulted in severe congestion in wait times and longer length of stays.

The (3,2) combination had the lowest average doctor wait (1.73) and the shortest LOS (95.31 min), indicating that this is the optimal solutions. The scatter plots and histograms explained in our video further emphasize this conclusion.

Model 3 improves the triage-prioritized framework by incorporating time-varying physician staffing. The capacity experiments show that bed supply is sufficient, while the alignment of physician capacity with demand is essential. Heavy daytime congestion under (1,4) split is eliminated by reallocating doctor into the day shift with (3,2) split. This came out as the most effective configuration for minimizing doctor wait and LOS. Overall, Model 3 makes a more realistic tool for staffing policy design.

3.4 Performance Metrics and Statistical Analysis

System performance is evaluated using average bed waiting time, average doctor waiting time, average length of stay, doctor utilization rate, and bed utilization rate.

4 Conclusion

4.1 Describe the major design decisions in your implementation

A major design decision was to make our model more realistic by moving beyond a simple FIFO baseline model as well as utilizing a Monte Carlo approach that does not fully incorporate the randomness and inherent variability of emergency department operations. Therefore, to address the limitations, we modeled the emergency department as a two-stage capacity-constrained system in which patients must first acquire a bed and then receive service from a doctor. Beds and doctors were modeled as separate constrained resources, with FIFO used for bed assignment and priority-based queuing for doctors based on triage level. Doctor capacity was further modeled to vary across day and night shifts while still holding total staffing constant, enabling us to evaluate how reallocating capacity over time impacts congestion and patient wait times.

4.2 What challenges did you encounter?

It was difficult to find data for the breakdown between waiting time and treatment time by triage level for emergency departments in the United States. As a result, we had to rely on Spanish emergency department data and assumed the distributions were transferable as triage processes are relatively similar in both countries. This lack of data problem persisted as we were looking for ways to expand our modeling into related topics like nurses, medication, or proximity to other major hospitals. Another challenge was handling the priority queue. We were able to combat this challenge by defining clear variable names and comments to keep track of logic.

4.3 What did you learn through implementing the project? Any major takeaways?

A key takeaway was that doctor availability, rather than bed capacity, was the dominant constraint in the system. It is likely that by similarly modeling a multitude of more constraints like number of nurses and inpatients hospitals can identify the exact values to operate at maximum efficiency. Although bed availability was sufficient, doctor utilization rates remained high, indicating a lack of readiness for any surges in throughput. These models showed the effetiveness of simulation modeling for identifying emergecy department bottlenecks given the assumptions we selected.

4.4 What are the next steps in the project if you were to continue working on it?

Our model does not account for seasonality or time of day peaks/drops. This was hard to incorporate as we only modeled 24 hours at a time. In the future, this project could be expanded by including longer simulation: 1 week, 1 month, or 1 year. This way the seasonality and surge trends can be included. In addition, our model oversimplifies with exponential distributions and ignoring other hospital resources. To improve this aspect we should incorporate more real world data to minimize oversimplification.

4.5 How do the software packages you used work? Are there other approaches, and what are the trade-offs?

Aside from generic python packages like pandas, numpy, and matplotlib our simulation did not require any other packages. While we did import heapq we did not end up utilizing it as it was not necessary due to the simplicty of our models. More complex packages like queuing-tool and heapq would be necessary for some of the next steps we suggested in the previous question.

4.6 What were the results of your experiments?

Across all three models, patient delay was primarily driven by doctor wait times, while bed wait time remained consistently at zero. This occurred because our bed capacity parameter was set sufficiently high so patients could secure beds immediately upon arrival. As a result, congestion accumulated in the priority-based doctor queue, leading to extended lengths-of-stay during peak periods, particularly for lower-acuity patients. Model 3's staffing sensitivity analysis revealed that holding total doctor capacity constant while reallocating physicians across day and night shifts improved performance more than increasing total staffing. This demonstrated that mismatches between demand and physician availability, rather than insufficient overall capacity, was the primary driver of congestion.

4.7 Do these results seem reasonable? Are they physically realistic?

The dominance of doctor wait time is consistent with real emergency department scenarios. However, the absence of bed wait time is not realistic, as beds are often a binding constraint during periods of high demand. The model also excludes several outside factors that would increase realism, including bed turnover delays, nurse staffing constraints, and diagnostic bottlenecks such as imaging and lab turnaround times. While the results are internally consistent, incorporating these elements and varying bed capacity would produce more realistic wait time and length of stay estimates.

4.8 What did you learn by running the experiments?

The experiments revealed strong sensitivity to doctor staffing levels, in which small changes in doctor allocation led to dramatic reductions in wait times. Visualization outputs showed doctor wait times spiking around shift changes. This shows the sensitivity to changing staff levels. Running 50 Monte Carlo replications also emphasized the importance of analyzing variability in our system, as confidence intervals revealed patterns that single-run simulations would miss.

4.9 How does your project fit into the larger context of the class and the IEOR field?

This project directly applies core concepts from INDENG 174, including Poisson arrival modeling, exponential service times, queueing, discrete-event simulation, and Monte Carlo simulation. The main purpose of this project was to optimize efficiecy for hospitals, increase balance for hospital staff, and improve patient care. Optimizing real world processes for ethical and efficiency reasons is at the core of industrial engineering and operations research. In addition, we were able to use a quantitative method to mitigate these issues which is at the focus of many other courses like linear programming and network flows and logistics network design. Lastly, the translation of data into actionable insights over time emphasizes the data driven aspect of industrial engineering and operations research.

5 Appendix: Simulation Code and Reproducibility

5.1 Programming Environment

The simulation was implemented and executed in the following environment:

- Language: Python 3
- Platform: Google Colab / Jupyter Notebook
- Operating mode: Single notebook with simulation, analysis, and plots
- Random seed: Explicitly controlled using `np.random.seed()` for reproducibility

5.2 Required Libraries

The implementation uses standard scientific Python libraries:

- numpy: for random variate generation and numerical computation
- heapq: (Python standard library) for priority-queue operations
- pandas: for structured data handling (optional in the final code)
- matplotlib: for plotting queue length trajectories and Monte Carlo distributions

5.3 Simulation Code

Please reference the attached pages at the end of this report for the final code that is implemented in Google Colab.

5.4 How to Run the Simulation

To reproduce the simulation and results:

1. Ensure Python 3 and the required libraries numpy, matplotlib, pandas, heap are installed.
2. Save the above code into a file, for example er_simulation.py, or into a Google Colab notebook cell.
3. (Optional) Set the random seed at the top of the script using np.random.seed(42) to ensure replicability.
4. Run the script:
    - Execute the cells sequentially in Google Colab / Jupyter or any Python IDE of your choice.
5. Each model will print:
    - Mean bed waiting time
    - Mean doctor waiting time
    - Mean bed and doctor utilization
    - Mean length of stay
    - Relevant plots
6. Additional plots may be added using matplotlib for visualization.