Emergency Department Staffing Optimization

1 Problem Statement

Emergency Departments (EDs) face highly variable patient arrival rates and a mix of patient acuity levels, leading to periods of overcrowding and patient delays. In this project, we seek to optimize ED staffing levels and policies to ensure timely care for patients while managing operational costs. The central problem is determining how to allocate and schedule medical staff in the ED in order to minimize patient waiting times and avoid excessive workload or idle time, given the stochastic nature of patient arrivals and service needs.

We aim to evaluate different staffing strategies under realistic ED conditions. Key performance metrics—such as patient waiting time, staff utilization, and staffing cost—will be quantified. These metrics will be combined into a single utility function that reflects overall performance. By comparing various staffing policies (including static schedules, dynamic adjustments, and threshold-based rules) via simulation, we will identify the policy that best balances efficiency (high utilization, low cost) with effectiveness (low patient wait times).

2 System Design and Components

The ED is modeled as a queueing system with time-varying patient arrivals and multiple patient categories (triage levels) requiring service from medical staff. We describe below the main components of the system and the assumptions for our simulation model.

2.1 Patient Arrival Process

Patient arrivals are modeled as a non-homogeneous Poisson process with a time-dependent arrival rate $\lambda(t)$. This means the rate of patient arrivals varies over the course of the day (for example, higher during daytime peak hours and lower overnight). The non-homogeneous Poisson assumption implies that the number of arrivals in any time interval $[t, t + \Delta t]$ is Poisson-distributed with mean $\int_t^{t+\Delta t} \lambda(u)\, du$, and arrivals in disjoint time intervals are independent. This stochastic model captures the random yet time-varying nature of ED patient inflow.

2.2 Triage Levels and Probabilities

Upon arrival, each patient is assigned a triage level based on severity, which determines their priority and expected service requirements. We consider a set of discrete triage categories (e.g., 1 = critical, 2 = urgent, 3 = non-urgent), with a fixed probability distribution for patient acuity. For example, a certain proportion of patients are critical (high acuity), some are urgent, and the rest are non-urgent. These probabilities reflect the case mix in the ED and sum to 1. Table **??** summarizes an example distribution of patients across three triage levels.

Triage Level	**Probability of Arrival**	**Mean Service Time (min)**
1 (Critical)	10%	60
2 (Urgent)	30%	30
3 (Non-urgent)	60%	15

Table 1: Example triage level distribution and service time parameters.

2.3 Service Times by Triage Level

Service times (i.e., the time a doctor or nurse spends treating a patient) are assumed to follow an exponential distribution for each triage category. The mean service time depends on the triage level, reflecting the greater complexity or resource needs of higher-acuity patients. As shown in Table **??**, for instance, a critical patient (Level 1) might have a mean service time of 60 minutes, whereas a non-urgent patient (Level 3) has a shorter mean of 15 minutes. We denote by μ_i the service rate for triage level i, so that the mean service time is $1/\mu_i$. Using exponential service times is a common assumption that provides memoryless service durations and simplifies the modeling, while still capturing the variability in treatment times.

2.4 Staffing Policy Descriptions

We consider several types of staffing policies for the ED, which define how the number of on-duty staff is determined or adjusted over time:

- **Static staffing:** A fixed number of staff (e.g. doctors or nurses) is scheduled for the entire period or for each shift, regardless of the current patient load. This policy is simple to implement but may be inefficient during slow periods or inadequate during surges, since it does not adjust to real-time demand.
- **Dynamic staffing:** The number of staff on duty varies according to a preset schedule or predicted demand pattern. For example, more staff can be scheduled during peak hours of the day and fewer staff during the night. This time-dependent staffing aims to better match the expected arrival

rate $\lambda(t)$, improving efficiency compared to static staffing. The dynamic schedule is determined in advance (e.g., based on historical arrival data or forecasts).

- **Threshold-based staffing:** An adaptive policy that adjusts staff in real time based on the current system state (such as the queue length or waiting time). In a threshold policy, additional staff are called in or activated when a certain threshold is exceeded (for instance, if the number of patients in the waiting room goes above Q or if the longest wait exceeds a set time). Likewise, staff may be allowed to leave or not replaced when the queue falls below a lower threshold. This reactive approach can ensure rapid response to surges, though it may be challenging to implement due to staff availability and short lead times.

Each of these policies will be modeled and tested in the simulation. Static staffing provides a baseline, dynamic staffing introduces proactive adjustments, and threshold-based staffing offers reactive control. The performance of these policies will be compared using the metrics described in the next section.

3 Performance Metrics and Utility Function

To evaluate staffing policies, we measure several performance metrics that capture different aspects of ED operation. We then combine these metrics into a single composite utility function that quantifies the overall effectiveness of a policy (with appropriate trade-offs).

3.1 Average Waiting Time

The *average wait time* $\bar{W}$ is the mean time that patients spend waiting in the queue before being seen by a medical provider. This metric is a key indicator of service quality and patient satisfaction in the ED. A lower average waiting time implies that patients are receiving quicker care. In our simulation, $\bar{W}$ will be estimated by tracking each patient's waiting time from arrival until service start, and averaging across all patients. Reducing $\bar{W}$ is usually a primary objective, but it must be balanced against resource utilization and cost.

3.2 Staff Utilization

Staff utilization is the fraction of time that the available staff are busy treating patients (as opposed to idle). We denote this metric by ρ. High utilization (near 100%) indicates that staff are almost constantly busy, which is efficient from a cost perspective but might risk staff burnout or long waits if demand exceeds capacity. Low utilization means staff are often idle, indicating a possible overstaffing.

We calculate utilization as the total busy time of all staff divided by the total time staff are available. For example, if N_s is the number of staff on duty

and B_i is the total time staff member i is busy during the simulation period of length T, then one can compute overall utilization as:

$$\rho = \frac{\sum_{i=1}^{N_s} B_i}{N_s \times T}\,, \tag{1}$$

where $N_s \times T$ is the total staff time available (e.g., N_s staff over T hours). In essence, ρ is the average fraction of time each staff member is working. The simulation will track staff busy time to empirically determine ρ for each staffing policy.

3.3 Staffing Cost

The *staffing cost* metric captures the resource expenditure on staffing. We assume a cost rate per staff (e.g., wage per hour for a doctor or nurse) denoted c_s. The total cost C for a given policy is proportional to the total staff-hours utilized. If $N(t)$ is the number of staff on duty at time t, the cost over a horizon T can be expressed as:

$$C = c_s \int_0^T N(t)\, dt\,, \tag{2}$$

which, in the simple case of a constant staff level, simplifies to $C = c_s \times N \times T$. In practice, this means that having more staff on duty or keeping staff on for longer hours increases the cost linearly. This metric discourages overstaffing by penalizing policies that use a large workforce for extended periods. Our simulation will compute C by accumulating the staffed hours times the cost rate.

3.4 Composite Utility Function

Because there is a trade-off between the above metrics (for example, we can reduce wait time by adding staff, but that increases cost and might lower utilization efficiency), we define a single *utility function* U to evaluate each policy's overall performance. This utility is a weighted linear combination of the key metrics, allowing a unified comparison. One convenient formulation is:

$$U = -\,w_1\,\bar{W} \;+\; w_2\,\rho \;-\; w_3\,C\,, \tag{3}$$

where $\bar{W}$ is the average waiting time, ρ is staff utilization, and C is the staffing cost. The weights w_1, w_2, and w_3 are positive coefficients that reflect the relative importance of each metric. In the formula, $\bar{W}$ and C are given negative weights because lower wait times and lower costs are desirable (so they contribute positively to utility when they decrease). Staff utilization ρ is given a positive weight since a higher utilization (up to a reasonable level) is preferred. By adjusting these weights, hospital administrators can prioritize what is most important in their context (for example, placing more weight on waiting time for patient satisfaction, or on cost if budget is a primary concern).

The goal is to choose a staffing policy (and associated parameters) that maximizes U. In effect, maximizing U yields an optimal trade-off among waiting time, utilization, and cost according to the specified weights.

4 Optimization Strategy

To identify the best staffing policy under the given model, we employ a simulation-based optimization strategy. Because the ED system is complex (with time-varying arrivals and random service times), analytic solutions for optimal staffing are intractable. Instead, we rely on simulation to evaluate the performance of different policies and search for an optimal or near-optimal solution.

4.1 Simulation-Based Evaluation

We use discrete-event simulation to model the ED operation for each candidate staffing policy. For a given policy (static, dynamic, or threshold-based with certain parameters), the simulation will replicate patient arrivals (according to $\lambda(t)$) and services, keeping track of queue lengths and events (arrivals, service start, service completion). The performance metrics (average wait $\bar{W}$, utilization ρ, cost C) are measured from the simulation output. By running the simulation for a sufficiently long period or multiple replications, we obtain reliable estimates of the metrics for that policy.

This simulation-based approach allows us to evaluate how a staffing policy would perform in practice, accounting for the randomness and time variability in the ED. The stochastic simulation acts as a "black-box" function that, given a policy and its parameters, produces the resulting performance metrics (and thus a utility U). While simulation provides estimates with some statistical noise, increasing the number of replications can make these estimates as accurate as needed.

4.2 Comparison of Policy Types

Using the simulation model, we will compare the different staffing policies (static, dynamic, threshold-based) outlined earlier. Each policy will be configured in a reasonable way for testing: for example, static staffing with a certain fixed number of providers, dynamic staffing with a certain schedule of staffing levels throughout the day, and threshold-based rules with chosen threshold values for adding or removing staff.

By analyzing simulation results, we can observe the strengths and weaknesses of each policy type. A static policy might be simpler but could lead to long waits during peak times or low utilization in off-peak times. A dynamic policy aligned to the expected demand curve should improve average waiting times and maintain higher utilization during busy hours, at the expense of slightly increased complexity in scheduling. A threshold-based policy can be very responsive to unexpected surges, potentially keeping waits low, but it

may be harder to implement due to the need for on-call staff and may result in fluctuating utilization.

We will compare the composite utility U for each policy type, as well as examine the individual metrics, to understand the trade-offs. This comparison will show which approach yields the best overall performance for the ED scenario modeled.

4.3 Advanced Optimization Strategies

Beyond comparing a few predetermined policies, we can explore systematic optimization of staffing decisions. Several optimization strategies can be applied in conjunction with the simulation model:

- **Grid search:** This approach entails discretizing the decision variables (e.g. number of staff in each shift, or threshold values) into a grid of possible values and exhaustively simulating each combination. For example, we might try static staffing levels of 5, 6, 7, 8 staff and see which yields the highest utility U. While simple, grid search can become computationally expensive if there are many decision variables or a fine grid, but it guarantees finding the best solution on that grid.

- **Black-box optimization:** Here we treat the simulation model as a function $f(\mathbf{x}) = U$ mapping policy parameters $\mathbf{x}$ (such as staffing levels at different times, or threshold values) to the utility outcome. We can apply general-purpose optimization algorithms that do not require gradient information. Examples include heuristic methods like genetic algorithms, simulated annealing, or Nelder–Mead simplex search. These methods iteratively propose new sets of parameters, run simulations, and use the observed U values to guide the search towards better policies. Black-box optimization can be more efficient than grid search in high-dimensional parameter spaces.

- **Surrogate modeling:** In this strategy, we first run simulations for a sample of different policy configurations and use that data to fit an approximate model (surrogate) of the utility function. For instance, a regression model or a machine learning method could predict U given the policy parameters. This surrogate is much faster to evaluate than the simulation. We can then optimize the surrogate model using standard optimization techniques to suggest an optimal policy, and finally validate that candidate policy with the actual simulation. Surrogate modeling (also known as metamodeling) can significantly reduce the computational cost of finding near-optimal solutions.

These advanced strategies are optional pathways to refine the search for the optimal staffing policy. Depending on the complexity of the problem and computational resources, we may choose a suitable approach. The end goal remains to identify a staffing configuration (possibly including how it changes

over time or in response to system state) that maximizes the utility U and thereby provides the best balance between patient service levels and operational efficiency in the ED.