Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

HYBRID MODELING AND SIMULATION FOR ENHANCING PATIENT ACCESS, SAFETY AND EXPERIENCE

Vishnunarayan Girishan Prabhu[1], Anupama Ramachandran[2], and Steven Alexander[2]

[1]School of Modeling, Simulation and Training at the University of Central Florida, Orlando, FL, USA
[2]Office of Patient Experience, Stanford Medicine Health Care, Newark, CA, USA

ABSTRACT

In recent years, hybrid modelling and simulation have become increasingly popular in healthcare for analyzing and improving systems such as patient flow, resource allocation, scheduling, and policy evaluation. These methods combine at least two simulation approaches discrete-event simulation, system dynamics, and agent-based modelling and may also integrate techniques from operations research and management sciences. Their ability to represent complex, dynamic systems has driven their adoption across various healthcare domains. This paper presents a case study of an emergency department (ED) where a hybrid framework combining forecasting models, hybrid simulation, and mixed-integer linear programming was used to optimize physician shift scheduling and improve patient flow and safety. The model outperformed current practices by reducing patient handoffs by 5.6% and decreasing patient time in the ED by 9.2%, without a budget increase. Finally, we propose incorporating reinforcement learning in future work to enable adaptive, data-driven decision-making and further enhance healthcare delivery performance.

1 INTRODUCTION

The Emergency Department (ED) is a vital component of the healthcare system, providing care for a wide range of conditions, from life-threatening emergencies to chronic and non-emergent issues. Under the Emergency Medical Treatment and Labor Act (EMTALA), the ED must screen and stabilize all patients regardless of their ability to pay, making it a primary access point for healthcare services (Laxmisan et al. 2007; McDonnell et al. 2013). Over the last several years, patient arrivals to EDs in the US have increased from 96.5 million annual visits in 1995 to 115.3 million in 2005 and 151 million in 2019 (Cairns et al. 2019; Centers for Disease Control and Prevention 2010). At the same time, the number of EDs in the US has decreased by over 15% in the last decade (Hsia et al. 2011). The ever-increasing patient volume and the decreasing number of EDs lead to a mismatch, predisposing EDs to crowding (Di Somma et al. 2015; George and Evridiki 2015; Kelen et al. 2021). The American College of Emergency Physicians (ACEP) defines ED crowding as the situation that "occurs when the identified need for emergency services exceeds available resources for patient care in the ED, hospital, or both" (American College of Emergency Physicians 2019). ED crowding is a global concern, often linked to suboptimal care, treatment delays, and increased risk of medical errors (Di Somma et al. 2015; Kulstad et al. 2010).

A few leading causes of ED crowding include high patient census (patient arrivals), inadequate resources (beds, medical devices, etc.), inadequate planning, and poor ED design (Morley et al. 2018; Moskop et al. 2009). Some of the most commonly adopted solutions to avoid ED crowding include expanding ED capacity, stopping boarding admitted patients in the ED, adding hallway beds, increasing on-call providers, and adding temporary resources (Derlet and Richards 2008). While these solutions are effective temporary fixes, they can often be costly and negatively impact patient safety and physician well-being. A recent study investigating ED crowding identified that access to future patient demands (arrivals

to ED) during shift scheduling and resource allocation can improve ED planning and potentially avoid crowding (Kelen et al. 2021). Most EDs, including our partner ED, build their clinician schedules about one quarter (3 months, 90 days) ahead. Hence, it is critical to have robust 90-day forecasts to assist ED administrators in planning clinician schedules to improve ED performance.

Researchers have used various forecasting methods to predict patient arrivals to the ED for different horizons (Aboagye-Sarfo et al. 2015; Batal et al. 2001; Carvalho-Silva et al. 2018; Choudhury and Urena 2020; Côté et al. 2013; Hertzum 2017; Jones et al. 2008; Kadri et al. 2014; Khaldi et al. 2019; Sun et al. 2009; Whitt and Zhang 2019; Xu et al. 2013; Zhang et al. 2022). Additionally, a few studies have focused on forecasting specific types of patient arrivals to the ED (primarily patients with respiratory diseases) (Becerra et al. 2020; Rosychuk et al. 2015). Various forecasting methods, such as Autoregressive Moving Average (ARMA), Vector ARMA (VARMA), Holt-Winters, linear and multiple linear regression, Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and machine learning models (ANNs and RNNs), have been widely applied to predict patient arrivals at hourly, daily, and monthly intervals. Prior research indicates that incorporating variables like weather, holidays, and local events can enhance forecast accuracy; however, these methods often rely on complex, expert-driven data preprocessing, limiting their practicality for routine use by ED administrators. A notable gap in existing studies is the lack of forecasting by Emergency Severity Index (ESI) levels, a standardized 1-to-5 scale used across North America that reflects patient severity and anticipated resource needs.(Hossein et al. 2013). Hence, from a resource allocation standpoint, including ESI levels in the patient forecasts is critical for better planning the ED resource allocation.

A variety of operations research approaches, including mathematical optimization models, queuing theory models, simulation modeling, and probabilistic models, have been used to address various challenges observed in the ED, including resource allocation (Ahsan et al. 2019; Connelly and Bair 2004; Elalouf and Wachtel 2021; Rais and Vianaa 2011). While simulation is popular for addressing operational issues, its outcome is a realization and not an optimal solution. However, by identifying a specific objective, researchers have developed mathematical models to identify optimal staffing levels, generate schedules, determine optimal bed or other resource requirements, etc. One study used a mixed-integer linear programming (MILP) model to minimize understaffing with respect to patient volumes, which resulted in significant improvements to median length of stay, door-to-provider time, and door-to-bed time (Sir et al. 2017). Researchers have also used a combination of simulation-optimization models to identify optimal solutions and test them in the simulation model to validate them (Ghanes et al. 2015). Most of the ED studies have focused on identifying solutions that reduce patient waiting time, reduce length of stay or improve ED throughput. While a few studies have used patient wait times as surrogates for patient safety, we introduce a new metric – handoffs, directly quantifying patient safety (Maughan et al. 2011).

Patient safety is a crucial part of the ED as continuous patient flow and interactions with multiple departments and providers make it prone to errors. Additionally, researchers have observed ED as one of the hospital departments with high error rates. Among the various factors that contribute to medical errors, handoffs - the transfer of a patient's care and responsibility from one provider to another have been identified as a major patient safety issue (Maughan et al. 2011; Venkatesh et al. 2015). Specifically, studies investigating ED handoffs observed that the vital signs were not communicated for approximately 75% of the patients, and errors were observed in about 60% of cases (Venkatesh et al. 2015). Hence, while developing the ED physician shift schedules, it is crucial to consider patient safety metrics such as handoffs.

Our literature review shows that researchers have used time series forecasting, mathematical models, and simulations to improve ED operations. However, to our knowledge, these studies have not combined these approaches and included patient severity in their forecast. Moreover, as noted earlier, few have considered a direct patient safety metric in the mathematical model. To address these research gaps, we first develop forecasting models for predicting the 90-day patient arrivals to the ED, including their ESI

level, which is then inputted into the mathematical model for developing the schedule. The objective of the mathematical model was to identify optimal shift schedules that minimize the combined cost of patient wait times, handoffs, and physician shifts, thus considering the patient flow, patient safety, and staffing budget to generate schedules. Further, to test the impact of the generated schedules on the ED performance, we used our validated simulation model (Girishan Prabhu et al. 2022).

2 DATA

Input data for developing the forecasting model included two specific data points: the time of the day and the ESI levels assigned to the patient presenting to the ED. However, for developing the optimization and models, other data points, including the number of beds, allowable physician shifts, patient arrivals, ESI level of the patients, patient time in the ED, and the number of interactions between physicians and patients, were gathered from the partner ED. Additionally, the research team included ED physicians working in the partner ED for guidance and addressing any other physician-dependent activities in the ED to be included in the model. The partner hospital, Prisma Health, is the largest healthcare provider in the state and serves as a tertiary referral center. The flagship academic ED is an Adult Level 1 Trauma Center seeing over 106,000 patients annually. Figure 1, below, represents the patient arrivals to the ED averaged for 2017-2019, used for our forecasting model.

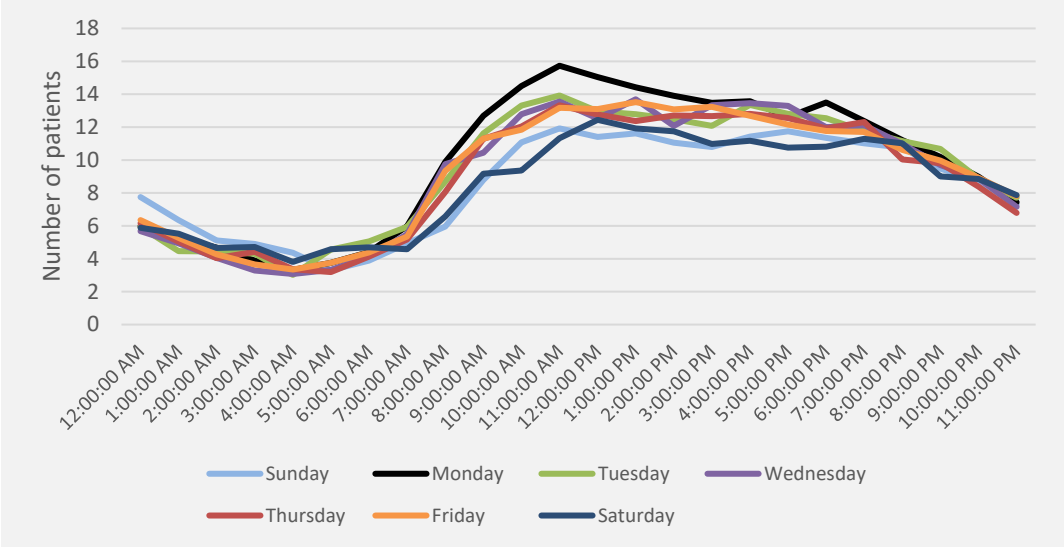

Figure 1: Patient arrivals to the ED.

One of the first noticeable patterns is the impact of the time of day on patient arrivals: arrivals are low during the early morning hours, gradually increase from 7:00 am, peak around noon, and remain steady until 7:00 pm. This trend aligns with findings from prior studies (Alvarez et al. 2009; Whitt and Zhang 2019). Another clear pattern is the difference between weekdays and weekends, with weekdays experiencing higher patient arrivals and Mondays having the highest volume. These observations highlight the need for long-term planning, as staffing requirements vary by day and hour. Regarding ESI levels, approximately 50% of arrivals are ESI-3 patients, followed by ESI-2 and ESI-4 at 25% and 20%, respectively. ESI-1 and ESI-5 patients each account for only 2–3% of arrivals. ESI-1 patients are critically unstable and require immediate intervention, while ESI-5 patients are the most stable and typically need minimal resources. Incorporating ESI levels into forecasting is essential since each level demands different resources.

For training the forecasting model, we used data from 2017, 2018, and the first six months of 2019. Validation and testing were performed during the last three months of 2019, which were divided into three-

month clusters. Rather than using the entire year of patient arrival data to generate physician schedules, we selected the cluster with the highest patient arrivals, specifically, the last three months, as the representative period for this study. Using the entire year's data was not feasible due to high variability, while relying solely on daily or weekly data would overlook operational factors such as staff leave or vacations that can affect patient flow. Clustering the data by quarters addressed these concerns. Additionally, based on input from ED physicians, we used pre-COVID-19 data to avoid the significant disruptions in patient arrivals seen during 2020. Another reason for using this specific period was to test the optimal schedule in our validated simulation model that used the same patient arrivals. However, both models were developed such that any patient arrivals can be used to generate a weekly schedule. Next, we introduce Table 1, which represents the time a patient spends in the ED based on their ESI levels.

Table 1: Patient time in the ED.

Severity	Bed to Disposition (mins)	Disposition to ED Departure (mins)	Total Time (mins)
ESI 1	115	121	236
ESI 2	186	86	272
ESI 3	175	54	229
ESI 4	90	24	114
ESI 5	107	15	122

As seen above, we split the data into two parts: "Bed to Disposition" and "Disposition to ED Departure." Bed to disposition represents the time a patient occupies an ED bed and is provided care by physicians and other medical providers, including performing tests, providing medicines, blood draws, etc. Although patients will be waiting in their beds during this period without receiving direct care, all these delays are due to waiting for their test results, medicines, etc. In general, this represents the period a patient first occupies a bed in the ED until the physicians make a disposition decision (admit, discharge, or transfer). The second part, "Disposition to Departure," is the period for which a patient occupies the ED bed from the time the physician makes a disposition decision until they are physically moved from the ED (discharged, admitted, or transferred). Hence, these are logistical delays where a patient can be either waiting until a bed is available in the hospital (admission) or waiting for transportation (discharged or transfer). While we primarily focus on the bed-to-departure time for this study, our model still accounts for delays before assigning a bed in the ED, where the patients wait in a waiting room until the beds are available, similar to an actual setting. As mentioned earlier, the entire bed-to-disposition time of a patient is not spent with a physician, as it includes other activities. Based on literature and discussions with ED physicians, we used between 15-30% of total time as the care time where a patient would be cared for by a physician (Füchtbauer et al. 2013). The percentages were assigned based on severity, such that the total time spent with an ESI-1 patient was the highest and that with an ESI-5 patient was the lowest. This approach was mainly used because of the lack of detailed visit-by-visit data available to support detailed modeling.

Further, to build a model representative of ED operations where a physician visits patients multiple times based on their severity (ESI-level), we split the care time into multiple smaller windows. Based on our past observational studies and discussion with ED faculty and physicians, on average, an ESI-1 patient was visited four times by a physician, ESI-2 and 3 were visited three times, and ESI-4 and 5 were visited two times (Girishan Prabhu et al. 2020). The physician's time with a patient for each visit was a constant time block of 15 minutes, as the MILP modeling approach considers time as a discrete block of events.

3 MODEL DEVELOPMENT

3.1 Forecasting Models

In this study, the moving average naïve model was used as a benchmark to compare against other forecasts. Based on the literature and data patterns, we developed ARIMA and SARIMA models, as they are well-

suited for time series forecasting, with SARIMA accounting for seasonality.(Carvalho-Silva et al. 2018; Choudhury and Urena 2020; Hertzum 2017; Kadri et al. 2014; Sun et al. 2009; Whitt and Zhang 2019). Additionally, we developed a Holt-Winters forecasting model as this approach can account for the level, trend, and seasonality components in the time-series data. Finally, we also developed two machine learning models: Extreme Gradient Boosting (XGBoost) and the Random Forest Regression model. Both are decision tree machine learning algorithms and require a supervised learning approach where each input requires an output pair within the training model for the model to learn and predict. However, the foundation of each algorithm is different, where Random Forest Regression uses a bagging technique, whereas XGBoost uses a boosting technique.

To prevent overfitting and properly tune hyperparameters, we employed blocked cross-validation. This method avoids data leakage and pattern memorization inherent in k-fold cross-validation for time-series by adding buffer zones: (i) between the training and validation folds and (ii) between the folds used at each iteration. Finally, to evaluate the performance of each model, we used Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percent Error (MAPE).

3.2 Optimization Model

We formulate the ED physician shift scheduling problem as an MILP model. The primary objective is to determine optimal physician staffing levels that minimize patient onboarding time, waiting time after ED admission, and patient handoffs, while accounting for staffing costs. To compare these factors uniformly, costs are expressed in dollar amounts. Before formalizing the model, we identified key ED operational activities to accurately replicate the partner ED. These include accounting for variable patient arrivals by ESI level, modeling multiple patient–physician interactions with minimum delays for secondary care (e.g., imaging, blood draws), ensuring continuity of care by assigning the same physician unless a shift ends (handoffs), and limiting physician shifts to 8 hours. Next, we define the notation used in the MILP model.

The model included four sets and corresponding indices as follows: (i) I represents the set of patient arrivals to the ED indexed by I, (ii) K represents the set of possible physicians that can be staffed for a day indexed by k, (iii)T represents the set of time slots considered for staff scheduling indexed by t and (iv) M represents the set of physician visits required by a patient indexed by m.

Here, set I includes all the unique patient arrivals to the ED for a week, which totals more than 1500. Set K consists of the unique physician identification number that can start an ED shift for a day, with an upper threshold of 25 physicians per day. Further, T represents timeslots for an entire week (which varies based on slot length). Finally, set M includes values from 1 through 4, representing the patient's interaction with a physician. Next, we introduce the parameters considered in the model. Most of the parameters represent various patient characteristics, such as severity, arrival time, physician visits, and fixed time slots. These fixed time slots should be excluded when calculating patient wait time, as these delays are inherent. One parameter also defines the ED bed capacity.

- α_i represents the time slot of arrival for patient i.
- β_i represents the severity level of patient i.
- γ_i represents the total number of visits required by patient i.
- w_i represents the total time slots for patient i that should not be included when calculating waiting cost. This is because there must be a minimum time between subsequent patient–provider visits to account for inherent delays, such as laboratory tests, blood draws, and other procedures.
- C represents the total bed capacity of the ED.

Finally, we introduce the decision variables in the model:

- U_{ik} $\begin{cases} 1, \text{ If patient i served by physician k} \\ 0, \text{ otherwise} \end{cases}$

- $Ystart_{kt}$ $\begin{cases} 1, \text{ If physician k starts their shift at time slot t} \\ 0, \text{ otherwise} \end{cases}$
- Y_{kt} $\begin{cases} 1, \text{ If physician k is available for service at time slot t} \\ 0, \text{ otherwise} \end{cases}$
- X_{iktm} $\begin{cases} 1, \text{ If patient i is served by physician k at time slot t for their visit m} \\ 0, \text{ otherwise} \end{cases}$

Minimize:

$$SC^* \sum_{kt} Ystart_{kt} + OC^* \sum_{ikt} t^*X_{ikt1} - \alpha_i + OC^*F^* \sum_{ikt} (t^*X_{ikt\gamma_i} - t^*X_{ikt1} - w_i) + HC^* \sum_{ik} U_{ik}$$

Subject to:

$$\sum_{kt} t * X_{ikt1} \geq \alpha_i \qquad \forall i \in I$$

$$\sum_{ktm} X_{iktm} = \gamma_i \qquad \forall i \in I$$

$$\sum_{km} X_{iktm} \leq 2 \quad \forall i \in I, \forall t \in T$$

$$\sum_{kt} X_{iktm} = 1 \quad \forall i \in I, \forall m \in M$$

$$\sum_{ikm} X_{iktm} \leq C \qquad \forall t \in T$$

$$\sum_{kt} t^*X_{iktm} \leq \sum_{kt} t^*X_{iktm+1} \qquad \forall i \in I$$

$$\sum_{mt} X_{iktm} \leq 4^*U_{ik} \qquad \forall i \in I, \forall k \in K$$

$$\sum_{im} X_{iktm} \leq 4^*Y_{kt} \qquad \forall k \in K, \forall t \in T$$

$$\sum_{t} Ystrt_{kt} \leq 1 \qquad \forall k \in K$$

$$\sum_{kt} Ystrt_{kt} \leq K$$

$$8*\mathrm{Ystrt}_{kt} \leq \sum_{q=t}^{\mathrm{Min}(168,\, t+7)} \mathrm{Y}_{kq} \quad \forall k \in K, \forall t \in T$$

$$\mathrm{U}_{ik}\ , \mathrm{Ystart}_{kt}\ , \mathrm{Y}_{kt}\ , \mathrm{X}_{iktm} \quad \in \{0, 1\}$$

In the formulation, the objective function minimizes the cost of staffing the ED physicians, handoffs, patient onboarding, and patient waiting time in the ED. The cost of staffing an ED physician (*SC*) was based using the national average rate for ED physicians, and the onboarding cost (*OC*) for patients based on their ESI level was derived from the literature (Salary.com 2021; Woodworth and Holmes 2020). However, because of the lack of data on the cost of patient waiting once admitted, we used a factor value (*F*) between 0 and 1 and multiplied it by the *OC* to calculate the waiting cost. Finally, for the handoff cost (HC), a high value of $1,000 was selected, as sensitivity analyses indicated that further increases beyond this threshold did not yield significant changes in the resulting schedules. This value effectively penalizes unnecessary handoffs, thereby minimizing their occurrence within the model

The first constraint ensures that a patient is served their first visit (m=1) only after their arrival at the ED. The second constraint ensures that the patient is provided with all their required visits before discharge. As mentioned earlier, each hour represents a time slot, but from observations and discussions with physicians, we assume that a physician can visit four patients in an hour. However, the same patient cannot be visited four times in an hour, as that is not realistic, as patients wait to get their tests, imaging, radiology, etc., completed. The third constraint ensures that a patient can be visited at most twice by a physician in an hour. The fourth constraint assures that each visit *m* for a patient cannot exceed 1, making sure that each visit is completed fully during a physician visit. The next constraint ensures that at any given time *t*, the patients served cannot exceed the ED bed capacity. As patients have multiple interactions with physicians during an ED stay, these visits must be ordered such that a later visit (m+1) follows the prior visit (m) in terms of time slot, and our sixth constraint ensures the visits are ordered. The next two constraints ensure that a patient can be visited a maximum of four times by a physician, and a physician can visit up to four patients during any given time slot (1-hour block). The next two constraints ensure that a physician starts their shift only once a day and that the total number of physicians staffed per day does not exceed the maximum number of possible physicians that can work for a day based on health system budget constraints. To ensure that a physician shift, once started, lasts for eight hours, we use the second-to-last constraint. Finally, the last constraint defines the variable types, which are all binary in this case.

3.3 Simulation Model

After developing the mathematical model to generate staffing schedules, the next step involved creating and validating a simulation model representative of the partner ED. We employed a novel hybrid simulation approach by integrating discrete event simulation with agent-based modeling, where both patients and physicians are modeled as individual agents possessing distinct attributes. This hybrid framework enabled the detailed simulation of actual patient arrivals, capturing features such as severity levels, arrival times, and care pathways. More importantly, it allowed us to realistically replicate physician behaviors and workflows, including shift start times, time spent at workstations for ordering tests and updating patient records, multiple patient visits, and the handoff process when shifts end. Such complex, dynamic physician activities would be difficult to accurately represent using traditional discrete event simulation alone, where physicians are typically modeled as generic resources rather than autonomous agents. This hybrid approach thus enhances the fidelity of the simulation, improving its utility for evaluating staffing policies and operational outcomes.

Figure 2 provides a high-level overview of patient flow and physician activities within a single ED pod. Patients arrive according to historical data and first undergo triage, where they are assigned an Emergency Severity Index (ESI) level based on observed distributions. If no ED beds are available, patients wait in the

waiting room, prioritized by their severity level. Physicians arrive at the pod according to their scheduled shift start times. Upon arrival, a physician proceeds to the physician's station. If another physician is leaving at the same time, their patients are transferred to the incoming physician, representing handoffs consistent with ED practice. After handoffs, the physician reviews patient charts at the station before visiting patients as needed. Further, whenever there are free beds in the ED, a physician, based on their workload, will sign up a patient from the waiting room and meet them in their bed (or room). To replicate the actual assignment process followed in the ED, physicians working in certain pods were restricted from taking high-severity patients, as a few pods do not have the equipment required to provide care for high-severity patients. After visiting a patient for the first time, a physician always returns to the station to update charts and order tests. The patient will have subsequent visits by the same physician based on ESI level, as observed in the ED. If the physician is ending their shift, the patient is handed off to another active physician. Furthermore, after a subsequent visit, there is a 40% probability that a physician will see another patient before returning to the station. Where historical data were unavailable, expert input was used to inform the model.

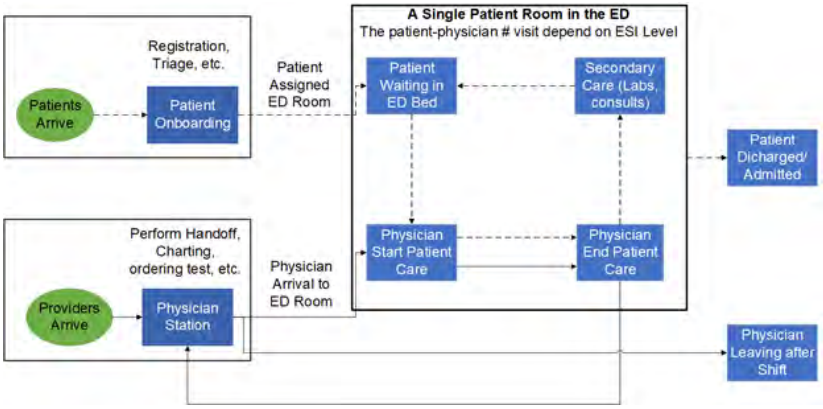

Figure 2: A high-level overview of patient and physician activities in a single ED pod.

After developing the simulation model representative of the partner health system's ED, we validated the model outputs against actual patient-level data. Specifically, patient time spent in the ED for each ESI level was used as the primary validation metric to ensure that the model outputs did not differ significantly from historical observations. Although patient-level data were compared, day-to-day variability was not explicitly used during validation. The model was simulated over a three-week schedule with an additional two-day warm-up period to reach steady state. A total of 60 replications were conducted to ensure that the margin of error for the time-in-ED metric was within ±10 minutes at a significance level (α) of 0.05. An independent t-test showed no statistically significant differences (p-value > 0.05) between the simulated and actual patient times. The average actual and simulated times in the ED were 236 and 218 minutes for ESI 1, 272 and 281 minutes for ESI 2, 229 and 216 minutes for ESI 3, 114 and 121 minutes for ESI 4, and both 122 minutes for ESI 5, with percent differences all below 7%.

4 RESULTS

As mentioned in the earlier section, the ED data averaged for 2017–2019 were used for our forecasting model. Upon developing each model, relevant parameters were tuned on the training data. For the traditional time series models, this included selecting appropriate seasonal and trend smoothing parameters (e.g., for Holt-Winters) and identifying the best-fitting p, d, q (and P, D, Q for SARIMA) terms. For the machine learning models (Random Forest and XGBoost), hyperparameters such as the number of trees, tree depth, learning rate, and minimum child weight were optimized through cross-validation. Following model development and tuning, the final models were used to forecast patient arrivals for a 90-day period. To clarify the temporal split: data from July 2017 through June 2019 were used for model training and

parameter tuning, and data from July 2019 through September 2019 were held out for validation and testing. Including two full years of data enabled the models to capture seasonal trends (e.g., daily, weekly, and yearly patterns), which is critical when forecasting over the next 90 days, where such seasonality may still influence arrival patterns.

Table 2 presents the performance metrics for each model's 90-day forecasts. Both machine learning models outperformed the naïve and traditional time series approaches. The Holt-Winters method performed better than ARIMA, likely due to its ability to explicitly capture seasonality. Compared to Holt-Winters, the SARIMA model showed slightly improved performance by combining autoregressive and seasonal differencing components. The most substantial improvements were observed with the machine learning models, where the MAPE was reduced by half relative to the traditional models. Among these, XGBoost outperformed Random Forest on all performance measures. A key observation is the high RMSE values across all approaches, likely due to outliers in daily arrivals. Even with significant fluctuations in patient demand, the machine learning forecasts remained robust across RMSE, MAE, and MAPE, with MAPE values near 5%, which are generally considered robust. To avoid over-reliance on MAPE alone, we interpret it alongside RMSE (e.g., RMSE of 16.6 against daily arrivals ranging from ~150 to 270), providing a more comprehensive view of forecasting accuracy.

After identifying XGBoost as the best-performing model, we examined ESI-level forecasts for the same 90-day period. Table 3 summarizes the performance metrics by ESI level. Notably, RMSE, MAE, and MAPE vary across levels: MAPE is higher for ESI 1 and 5 and lowest for ESI 3, while RMSE and MAE exhibit the opposite trend. This illustrates a known limitation of MAPE, where it disproportionately penalizes smaller values since it is a percentage-based metric. This reinforces the importance of using multiple metrics (RMSE, MAE, and MAPE) together to assess model performance comprehensively, rather than relying solely on MAPE where forecasted volumes are small.

Table 2: Model performance for the 90-day forecast.

Model	RMSE	MAE	MAPE
MA	30.1	23.6	14.2%
ARIMA	27.2	21.6	10.6%
Holt-Winters	26.8	19.8	10.0%
SARIMA	25.6	19.2	9.9%
Random Forest	17.4	14.6	6.4%
XGBoost	16.6	14.1	5.9%

Table 3: XGBoost ESI level 90-day forecast.

ESI	RMSE	MAE	MAPE
ESI 1	3.1	2.8	38.0%
ESI 2	8.9	7.0	12.5%
ESI 3	13.4	10.1	8.4%
ESI 4	6.0	5.1	15.5%
ESI 5	1.9	1.4	46.1%

The forecasted patient arrivals, along with other data such as the number of available ED beds, operational policies, physician shift constraints, and required provider–patient interaction times, were input into the mathematical model to identify staffing schedules that minimize patient handoffs, physician shifts, and patient wait times while staying within the staffing budget. Table 4 below presents the physician shift start times for a representative week in the study period, based on the forecasted patient arrivals for that specific week.

Table 4: Physician shift start times for a representative week.

Time	00	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Current	0	0	0	0	0	0	0	22	4	20	5	0	2	0	0	21	4	20	0	0	0	0	15	21
New	0	0	0	0	0	0	0	14	14	14	7	7	0	7	7	14	21	0	0	0	0	0	0	21

The table compares the baseline (current) scheduling policy to the new schedules generated by the mathematical model. The numbers indicate the total number of physician shifts starting at each hour of the day, aggregated over the entire week. While the operational policies restrict shift start times to certain hours (e.g., physicians cannot begin shifts at midnight, 2:00 am, etc.), the optimized schedule demonstrates a more staggered approach within the allowed time windows. Unlike the current policy, which tends to cluster

shift starts at specific times such as 7:00 or 9:00, the model-generated schedule distributes shift start times more evenly throughout the day to better match fluctuating patient demand and reduce handoffs, wait times, and staffing inefficiencies. Finally, in terms of total staffed hours for the week, there were no significant differences between the new and baseline schedules. After generating the new schedule, we compared it to the current (baseline) policy using the validated simulation model. Specifically, two key ED performance metrics were used: the number of handoffs and patient time spent in the ED. The first metric (number of handoffs) reflects patient safety, while patient time in the ED serves as an indicator of patient flow. Both policies were simulated over a representative three-week schedule, with replications performed until the margin of error for patient time in the ED was within ±10 minutes at a significance level ($\alpha = 0.05$).

Table 5: Simulation model results.

Policy	Weekly Throughput	# handoffs per day	Time in the ED (mins)	Change in hours/week (FTEs)
Baseline	1505	93	213±4.6	0
New	1503	84	201±5.9	+6 (+0.2 FTEs)

Based on the results presented in Table 5, an independent t-test showed that both handoffs per day and patient time in the ED differed significantly between the two policies (p-value < 0.05), with the new policy demonstrating improved performance. Compared to the baseline, the new policy reduced patient time in the ED by 5.6% and decreased handoffs by 9.6%, with only a slight, non-significant increase in full-time equivalents (FTEs) (p-value = 0.21).

5 CONCLUSIONS

Protecting the ED from crowding remains one of the highest public health priorities to ensure timely patient care and maintain patient safety. Although many EDs across the US develop advanced plans to prevent overcrowding, studies have shown that most still rely heavily on short-term, reactive measures. While ad hoc actions are sometimes necessary in response to unforeseen events such as evacuations or natural disasters, they are often a consequence of insufficient short- and long-term operational planning. A critical input for robust ED planning is an accurate forecast of future patient volumes. Over the past several decades, numerous studies have proposed various methods to forecast patient arrivals, generally achieving acceptable performance. However, most prior research has focused on predicting total daily patient arrivals, with only a few recent studies extending this to hourly forecasts (Choudhury and Urena 2020; Zhang et al. 2022). Notably, none of these studies have incorporated ESI levels into their forecasts, an important omission given that ESI levels directly influence staffing needs and resource allocation.

This research developed and evaluated both traditional time-series models and machine learning approaches to forecast long-term (90 days ahead) patient arrivals to the partner ED, including each patient's ESI level. Among the models tested, the XGBoost algorithm produced the most accurate forecasts, with a MAPE of 5.9%, outperforming results reported in prior studies. In addition, ESI-level forecasts yielded a maximum RMSE of 13.4, which is encouraging given the limited input variables and the extended forecast horizon. These forecasts were then integrated into a mathematical model to determine optimal physician staffing schedules that minimize onboarding time, waiting time after ED admission, and patient handoffs while staying within staffing constraints. The resulting schedule, when tested in the validated simulation model representative of the partner ED, reduced patient time in the ED by 5.6% and handoffs by 9.6%, with only a non-significant increase in FTE requirements.

Future research will aim to refine the forecasting model by incorporating additional simple parameters that can be readily extracted from the EHR to further improve predictive accuracy. A hierarchical forecasting approach with an embedded optimization function may also enhance ESI-level predictions. Finally, one limitation of the current mathematical and simulation models is the lack of explicit representation of ancillary resources such as laboratory services, consults, and nursing staff as separate

resources. Although these factors are indirectly captured through delays, future work will expand the model to explicitly include these resources and processes to more accurately reflect real-world ED operations.

ACKNOWLEDGEMENTS

We would like to thank Kevin Taaffe, Ronald Pirrallo, and William Jackson from Clemson Industrial Engineering and Prisma Health Emergency Department for their invaluable support and expertise.

REFERENCES

Aboagye-Sarfo, P., Q. Mai, F. M. Sanfilippo, D. B. Preen, L. M. Stewart, and D. M. Fatovich. 2015. "A Comparison of Multivariate and Univariate Time Series Approaches to Modelling and Forecasting Emergency Department Demand in Western Australia". *Journal of Biomedical Informatics* 57(1):62–73.

Ahsan, K. B., M. R. Alam, D. G. Morel, and M. A. Karim. 2019. "Emergency Department Resource Optimisation for Improved Performance: A Review". *Journal of Industrial Engineering International* 15(1):253–266.

Alvarez, R., G. A. Sandoval, S. Quijada, and A. D. Brown. 2009. "A Simulation Study to Analyze the Impact of Different Emergency Physician Shift Structures in an Emergency Department". *Proceedings of the 35th International Conference on Operational Research Applied to Health Services*, July 12th -17th, Leuven, Belgium, 900-902.

American College of Emergency Physicians. 2019. "Crowding". Policy Statement, American College of Emergency Physicians, Irving, Texas.

Batal, H., J. Tench, S. McMillan, J. Adams, and P. S. Mehler. 2001. "Predicting Patient Visits to an Urgent Care Clinic using Calendar Variables". *Academic Emergency Medicine* 8(1):48–53.

Becerra, M., A. Jerez, B. Aballay, H. O. Garcés, and A. Fuentes. 2020. "Forecasting Emergency Admissions due to Respiratory Diseases in High Variability Scenarios using Time Series: A Case Study in Chile". *Science of the Total Environment* 7(6):968-978.

Cairns, C., J. J. Ashman, and K. Kang. 2019. "Emergency Department Visit Rates by Selected Characteristics: United States". *NCHS data brief* 1(434): 1–8.

Carvalho-Silva, M., M. T. T. Monteiro, F. de Sá-Soares, and S. Dória-Nóbrega. 2018. "Assessment of Forecasting Models for Patients Arrival at Emergency Department". *Operations Research for Health Care* 18(4):112–118.

Centers for Disease Control and Prevention. 2010. NCHS Pressroom - Fact Sheet - Emergency Department Visits. https://www.cdc.gov/nchs/pressroom/04facts/emergencydept.htm, accessed 8th April 2023.

Choudhury, A., and E. Urena. 2020. "Forecasting Hourly Emergency Department Arrival using Time Series Analysis". *British Journal of Health Care Management* 26(1):34–43.

Connelly, L. G., and A. E. Bair. 2004. "Discrete Event Simulation of Emergency Department Activity: A Platform for System-level Operations Research". *Academic Emergency Medicine* 11(11):1177–1185.

Côté, M. J., M. A. Smith, D. R. Eitel, and E. Akçali. 2013. "Forecasting Emergency Department Arrivals: A Tutorial for Emergency Department Directors". *Hospital Topics* 91(1):9–19.

Derlet, R. W., and J. R. Richards. 2008. "Ten Solutions for Emergency Department Crowding". *Western Journal of Emergency Medicine* 9(1):24-36.

Di Somma, S., L. Paladino, L. Vaughan, I. Lalle, L. Magrini, and M. Magnanti. 2015. "Overcrowding in Emergency Eepartment: An International Issue". *Internal and Emergency Medicine* 10(2):171–175.

Elalouf, A., and G. Wachtel. 2021. "Queueing Problems in Emergency Departments: A Review of Practical Approaches and Research Methodologies". *Operations Research Forum* 3(1):1–46.

Füchtbauer, L. M., B. Nørgaard, and C. B. Mogensen. 2013. "Emergency Department Physicians Spend only 25% of their Working Time on Direct Patient Care". *Danish Medical Journal* 60(1):120-128.

George, F., and K. Evridiki. 2015. "The Effect of Emergency Department Crowding on Patient Outcomes Results". *Health Science Journal* 9(1):1–6.

Ghanes, K., O. Jouini, A. Diakogiannis, M. Wargon, Z. Jemai, R. Hellmann, V. Thomas, and G. Koole. 2015. "Simulation-based Optimization of Staffing Levels in an Emergency Department". *SIMULATION* 91(10):942–953.

Girishan Prabhu, V., K. Taaffe, R. Pirrallo, and D. Shvorin. 2020. "Stress and Burnout among Attending and Resident Physicians in the ED: A Comparative Study". *IISE Transactions on Healthcare Systems Engineering* 11(1):1–19.

Girishan Prabhu, V., K. Taaffe, R. G. Pirrallo, W. Jackson, and M. Ramsay. 2022. "Overlapping Shifts to Improve Patient Safety and Patient Flow in Emergency Departments". *SIMULATION* 98(11):961–978.

Hertzum, M. 2017. "Forecasting Hourly Patient Visits in the Emergency Department to Counteract Crowding". *The Ergonomics Open Journal* 10(1):1–13.

Hossein, A., J. Rouhi, S. Sardashti, A. Taghizadieh, H. Soleimanpour, and M. Barzegar. 2013. "Emergency Severity Index (ESI): A Triage Tool for Emergency Department". *International Journal of Emergency Medicine* 12(2):92-106.

Hsia, R. Y., A. L. Kellermann, and Y. C. Shen. 2011. "Factors Associated with Closures of Emergency Eepartments in the United States". *Journal of the American Medical Association* 305(19):1978–1985.

Jones, S. S., A. Thomas, R. S. Evans, S. J. Welch, P. J. Haug, and G. L. Snow. 2008. "Forecasting Daily Patient Volumes in the Emergency Department". *Academic Emergency Medicine* 15(2):159–170.

Kadri, F., F. Harrou, S. Chaabane, and C. Tahon. 2014. Time Series Modelling and Forecasting of Emergency Department Overcrowding. *Journal of Medical Systems* 38(9):1–20.

Kelen, G. D., R. Wolfe, G. D'onofrio, A. M. Mills, D. Diercks, S. A. Stern, M. C. Wadman, and P. E. Sokolove. 2021. "Emergency Department Crowding: The Canary in the Health Care System". *New England Journal of Medicine Catalyst*. 2(5):1–26.

Khaldi, R., A. El Afia, and R. Chiheb. 2019. "Forecasting of Weekly Patient Visits to Emergency Department: Real Case Study". *Procedia Computer Science* 14(8):532–541.

Kulstad, E. B., R. Sikka, R. T. Sweis, K. M. Kelley, and K. H. Rzechula. 2010. "ED Overcrowding is Associated with an Increased Frequency of Medication Errors". *American Journal of Emergency Medicine* 28(3):304–309.

Laxmisan, A., F. Hakimzada, O. R. Sayan, R. A. Green, J. Zhang, and V. L. Patel. 2007. "The Multitasking Clinician: Decision-making and Cognitive Demand during Team Handoffs in Emergency Care". *International Journal of Medical Informatics* 76(1):801–11.

Maughan, B. C., L. Lei, and R. K. Cydulka. 2011. "ED handoffs: Observed Practices and Communication Errors". *American Journal of Emergency Medicine* 29(5):502–511.

McDonnell, W. M., C. A. Gee, N. Mecham, J. Dahl-Olsen, and E. Guenther. 2013. "Does the Emergency Medical Treatment and Labor Act Affect Emergency Department Use?". *The Journal of Emergency Medicine* 44(1):209–216.

Morley, C., M. Unwin, G. M. Peterson, J. Stankovich, and L. Kinsman. 2018. "Emergency Department Crowding: A Systematic Review of Causes, Consequences and Solutions". *PLoS ONE* 13(8):1–42.

Moskop, J. C., D. P. Sklar, J. M. Geiderman, R. M. Schears, and K. J. Bookman. 2009. "Emergency Department Crowding, Concept, Causes, and Moral Consequences". *Annals of Emergency Medicine* 53(5):605–611.

Rais, A., and A. Vianaa. 2011. "Operations Research in Healthcare: A survey". *Intl Transactions in Operational Research* 18(1):1–31.

Rosychuk, R. J., E. Youngson, and B. H. Rowe. 2015. "Presentations to Alberta Emergency Departments for Asthma: A Time Series Analysis". *Academic Emergency Medicine* 22(8):942–949.

Salary.com. 2021. Physician - Emergency Room Salary in the United States. www.salary.com/research/salary/benchmark/er-doctor-salary, accessed 10th April 2023.

Sir, M. Y., D. Nestler, T. Hellmich, D. Das, M. J. Laughlin, M. C. Dohlman, and K. Pasupathy. 2017. "Optimization of Multidisciplinary Staffing Improves Patient Experiences at the Mayo Clinic". *INFORMS Journal on Applied Analytics* 47(5):425–441.

Sun, Y., B. H. Heng, Y. T. Seow, and E. Seow. 2009. "Forecasting Daily Attendances at an Emergency Department to Aid Resource Planning". *BMC Emergency Medicine* 9(1):1–9.

Venkatesh, A. K., D. Curley, Y. Chang, and S. W. Liu. 2015. "Communication of Vital Signs at Emergency Department Handoff: Opportunities for Improvement". *Annals of Emergency Medicine* 66(2):125–130.

Whitt, W., and X. Zhang. 2019. "Forecasting Arrivals and Occupancy Levels in an Emergency Department". *Operations Research for Health Care* 21(2):1–18.

Woodworth, L., and J. F. Holmes. 2020. "Just a Minute: The Effect of Emergency Department Wait Time on the Cost of Care. Economic". *Inquiry* 58(2):698–716.

Xu, M., T. C. Wong, and K. S. Chin. 2013. "Modeling Daily Patient Arrivals at Emergency Department and Quantifying the Relative Importance of Contributing Variables using Artificial Neural Network". *Decision Support Systems* 54(3):1488–1498.

Zhang, Y., J. Zhang, M. Tao, J. Shu, and D. Zhu. 2022. "Forecasting Patient Arrivals at Emergency Department using Calendar and Meteorological Information". *Applied Intelligence* 52(10):11232–11243.

AUTHOR BIOGRAPHIES

VISHNUNARAYAN GIRISHAN PRABHU is an Assistant Professor in the School of Modeling, Simulation and Training at the University of Central Florida. His research interests include developing mathematical models, simulation models and machine learning models to improve healthcare operations. His email address is vishnunarayan.girishanprabhu@ucf.edu and his website is https://www.cecs.ucf.edu/smst/person/vishnu-prabhu/

ANUPAMA RAMACHANDRAN is the Senior Director – PMO, Training and Knowledgebase in Enterprise Contact Center, Office of Patient Experience at Stanford Medicine Health Care. Her interests are optimizing healthcare operations, patient flow, care coordination and digital transformation for improving operations. Her email address is aramachandra@stanfordhealtcare.org

STEVEN ALEXANDER is the Vice President, Enterprise Contact Center Strategy & Operations at Stanford Medicine Health Care. His research interests include operations research applications in healthcare for improving patient experience. His email address is SAlexander@stanfordhealthcare.org.