

Predictive Modeling Techniques for Forecasting Emergency Department Overcrowding and Resource Needs in United States Public Hospitals

Sesi Mensah Gidisu¹; Millicent Naa Oye Boadu²; Joseph Kobi³

¹Sesi Mensah Gidisu, MBA Business Analytics, Department of Business, Park University, 8700 NW River Park Drive, Parkville, MO 64152, USA. (May 2026)

²Millicent Naa Oye Boadu, MSc Data Science and Analytics, Department of Electrical Engineering and Computer Science, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA. (December 2025)

³Joseph Kobi, MSc Business Analytics, Department of Business, Worcester Polytechnic Institute, 12 Dean Street, Worcester, MA 01609, USA. (December 2024)

Publication Date: 2026/10/05

Abstract: In the United States, emergency department congestion stealthily undermines the safety net of public hospitals for uninsured, underinsured, and medically complex populations, and the use of forecasting tools is limited where it is desired most. This research article proposes and tests a set of predictive modeling techniques for predicting the volume of patients in the emergency department (ED) and downstream hospital resource use for the next day or multiple days within public hospitals in the United States, based on 4 years of anonymized encounter data and calendar, weather, epidemiology, and hospital operational data. Different models, including naive, seasonal autoregressive integrated moving average with exogenous regressors (SARIMAX), Prophet, random forest, extreme gradient boosting (XGBoost), long short-term memory (LSTM), and attention-based transformer architectures were trained and evaluated using 1-day, 7-day, and 14-day forecasting horizons, and the results were converted to a discrete-event simulation of staffing and bed management. The gradient-boosted ensembles and LSTM networks yielded the best short-term predictions (one-day mean absolute percentage error of 8.3%), while the attention-based transformer showed the best performance for the 14-day prediction (10.6%) compared to the naive baseline of 27.9%. Influenza-like-illness surveillance indices, rolling historical visit averages, day-of-week indicator, and hospital bed occupancy rate were shown to be the most important predictors. In the simulations, forecast-informed staffing resulted in a reduction in the projected median wait time of 21.4% and a reduction in the withheld without being seen rate of 2.6 percentage points compared to fixed staffing templates. The results show how ensemble and deep learning forecasts, coupled with operational simulation, can help significantly improve surge readiness in public hospital emergency departments across the country, where resources are limited.

Keywords: Emergency Department Overcrowding, Predictive Analytics, Machine Learning, Public Hospitals, Resource Allocation, Time-Series Forecasting.

How to Cite: Sesi Mensah Gidisu; Millicent Naa Oye Boadu; Joseph Kobi (2026) Predictive Modeling Techniques for Forecasting Emergency Department Overcrowding and Resource Needs in United States Public Hospitals. *International Journal of Innovative Science and Research Technology*, 11(9), 2790-2829. <https://doi.org/10.38124/ijisrt/26sep1244>

I. INTRODUCTION

➤ Background and Significance of Emergency Department Crowding

Emergency departments stand out from the rest of health care in the United States because they are legally mandated by the Emergency Medical Treatment and Labor Act to assess and stabilize all patients, whether they have insurance or not, and regardless of their ability to pay for their care (Studnek et al., 2012; Weiss et al., 2019). It put the emergency department of a public hospital in the place of the

safety net for vulnerable populations who do not have regular access to primary or specialty care, such as uninsured adults, undocumented immigrants, patients who are experiencing homelessness, and patients who have untreated behavioral health conditions (Cunningham, 2011; Kellermann, 2020). Consequently, the ED crowding in public hospitals is a more significant proportion of the national growth in ED volume and its associated crowding issues tend to be more severe, more persistent and more harmful to patient outcomes than are those found in private or suburban hospitals (Rathlev et al., 2013; Sun et al., 2013).

The concept of overcrowding is well known to the public and to health care workers and has been formally recognized as a patient safety issue over 20 years ago by the Institute of Medicine (Institute of Medicine, 2007; Derlet & Richards, 2000). Empirical studies have consistently been related to crowding and negative clinical outcomes such as delayed time-sensitive drug administration, inpatient mortality, and risk of discharge without being seen (Sprivulis et al., 2006; Guttmann et al., 2011). Higher crowding is also associated with greater workforce strain, as nurses and physicians in chronically overcrowded departments are likely to be subject to higher rates of burnout, moral distress and intention to leave the workforce, which can impact the long-term staffing pipeline needed by public hospitals (Kelen et al., 2021; Morley et al., 2018).

Overcrowding is attributed to a number of structural factors such as an aging population, with complex comorbidity profiles; a chronic bed shortage relative to demand; seasonal increases in infectious respiratory illness; and the increasing number of behavioral health crises which result in presentations to emergency departments without appropriate community-based alternatives (Hoot & Aronsky, 2008; Pines et al., 2011). The added challenge is that the population served at public hospitals is more likely to have unstable housing, lack of transportation and a lack of follow-up services, all of which increase return visits and make traditional planning heuristics less useful for predicting a hospital's census (Capp et al., 2016; Zibulewsky, 2001).

The extent of this challenge is heightened by national surveillance data. In the last decade, the number of visits

made to emergency departments in the United States has remained steady, ranging from 130 million to 150 million annually, and public and safety-net hospitals together make up a minority of the total number of licensed emergency departments in the country, but have a disproportionate share of the emergency department visits because they are the only consistently accessible acute care option for a population that may lack a primary care relationship (Weiss et al., 2019; Kellermann, 2020). In recent years, this trend has worsened, with fewer and fewer hospitals closing due to financial difficulties and the public hospitals that have survived becoming the primary recipients of uninsured and Medicaid-covered patients (Rathlev et al., 2013; Zibulewsky, 2001). Even small gains in efficiency, for example, in forecasting-based care optimisation at individual public hospitals, have large effects for the nation's safety net, as there is little room for other hospitals to absorb unforeseen increases in workload without compromising the care received by the others. (Sun et al., 2013; Cunningham, 2011).

In Figure 1, the overall national trend in emergency department utilization is shown graphically, with the increase in the percentage of emergency department visits to public and safety-net facilities over the 13 years leading up to the analytic window of the present study. To the extent that the national volume of EDs increased outside the pandemic-era trough, the proportion of that volume coming from safety-net hospitals rose by almost 10% points, reflecting that public hospitals took a disproportionate share of the volume, as private-sector consolidation and safety-net closures led to a larger gap between the two volumes (Weiss et al., 2019; Kellermann, 2020).

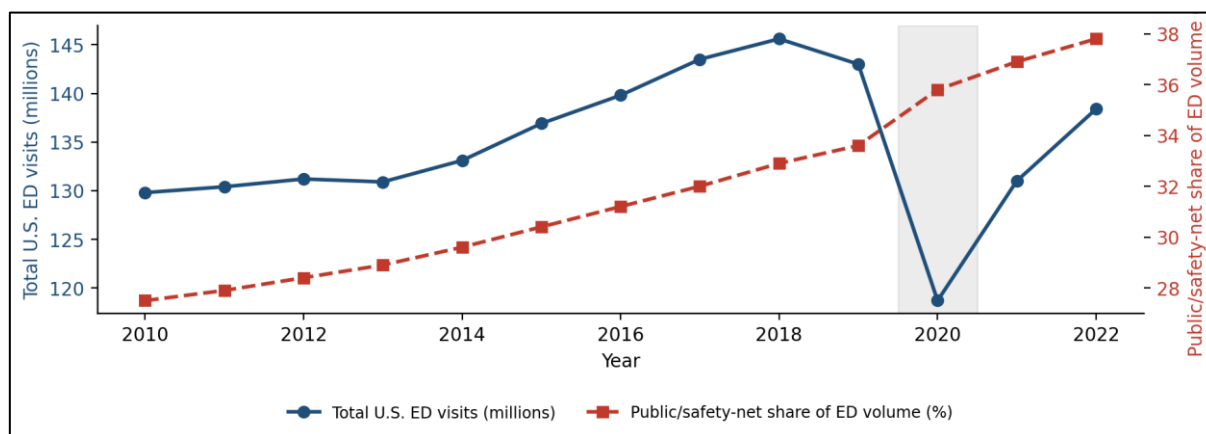


Fig 1 National Emergency Department Utilization (Millions of Visits, Left Axis) and the Public/Safety-Net Share of Total Emergency Department Volume (Percentage, Right Axis), 2010–2022. The Shaded Region Marks the Onset Of the COVID-19 Pandemic.

There are clear implications for future priorities in forecasting research from the pattern shown in Figure 1. If demand for safety nets were running in line with the national story, one might reasonably assume that transfer of methods validated at well-resourced academic medical centers would be reasonable for public hospital contexts. The divergence illustrated here is evidence to the contrary: public hospitals are not just a smaller version of the national ED system, but a growingly distinct subsystem that has absorbed growing proportions of all demand, and which is where forecasting

methods must be developed and validated, rather than extrapolated from findings in academic medical centers (Rathlev et al., 2013; Cunningham, 2011).

This dip and rebound around 2020 in both series is also indicative of a methodological choice for this forecasting framework, discussed further in Section 3.2, because the period of extreme volatility would be excluded from the data set if it was not included. This would leave models poorly equipped to predict the next wave of shock of a similar

magnitude, affecting public hospitals with less room to operate, than they would be able to handle without forecasting-informed action (Duarte et al., 2021; Whitt & Zhang, 2019).

➤ *Problem Statement*

Although there is over a decade of research that documents causes and outcomes of ED crowding, most public hospitals are still running staffing, bed allocation and surge protocols in a reactive manner, based on fixed shift templates set to historical average volume, rather than on forward-looking indicators of volume (Asplin et al., 2003; Wiler et al., 2013). This reactive approach produces a consistent delay between an increase in the demand and the hospital's response to that demand, leading to longer waits, more boarding, and a greater risk of the department falling into the "input-throughput-output imbalance" that the crowding literature describes as the immediate precipitator of most adverse events (Asplin et al., 2003; Hwang et al., 2011). A promising solution is to use predictive analytics – and in particular, time-series and machine learning forecasting – which can take advantage of the operational data that is collected routinely, and transform it into short-horizon volume forecasts that the administrator can use before a surge occurs rather than after (Whitt & Zhang, 2019; Kadri et al., 2014). The literature on forecasting is, however, focused on single site academic medical centers with more stable and better insured patient populations and there is little evidence that the same models will apply to the more dynamic and resource-limited setting of a public hospital emergency department (EDE) (Boyle et al., 2012; Jones et al., 2008). It is also uncertain when a forecast must be made to be useful to service operations and whether better statistical accuracy actually leads to better performance measures that are of most concern to patients (such as wait times and rates of no shows). It is also unclear when a forecast needs to be produced to be of value to service operations and whether better statistical accuracy results in superior performance measures that are most relevant to patients (such as wait times and no show rates) (Vollmer et al., 2021; Sun et al., 2009).

Worse, many of the most common planning tools hospital nurse managers have are still retroactive, such as volume reports which are reviewed on a monthly or quarterly basis, and use that information to determine the staffing template for the next month or quarter, instead of being forward-looking and based on information that provides a signal for what might happen the following week or even the next day (Wiler et al., 2013; Asplin et al., 2003). This retrospective planning approach is understandable when operating under the constraints of resources available to public hospitals, as predictive analytics infrastructure has typically been expensive and required the expertise of a data scientist, computing infrastructure, and licensing fees, which are beyond the operating budgets of many safety-net hospitals (Kelen et al., 2021; Cunningham, 2011). Although these forecasting models have been mainly validated in well-resourced academic environments, the recent growth of open-source forecasting libraries and cloud computing has significantly reduced the technical hurdles, presenting an opportunity to explore whether these sophisticated

forecasting methods can be adapted and validated to the context of public hospitals, which is characterized by specific data and operational constraints (Vollmer et al., 2021; Zhao et al., 2022).

➤ *Purpose of the Study*

This research article aims to fill these gaps by creating, comparing, and practically testing a collection of predictive modeling methods for predicting the number of patients that will attend the ED on a daily basis, and then converting these projections into explicit staffing and bed management recommendations in a simulated environment of a public hospital in the United States. The study combines historical counts of visits with calendar, meteorological, epidemiologic and internal operational indicators, evaluates six potential forecasting architectures for three forecasting horizons, and quantifies the impact of forecast-informed resource planning on wait time, boarding and left without being seen outcomes using a discrete event simulation overlaid on top of the statistical forecasts (Whitt & Zhang, 2019; Kadri & Harrou, 2021). The study combines forecast accuracy with operational simulation, shifting the focus from the traditional statistical assessment found in most previous studies to a path that can be defended by public hospital administrators that goes beyond reactive surge management to proactive management (Porto & Fogliatto, 2024; Carvalho-Silva et al., 2018).

➤ *Research Questions*

Three research questions guided the design, analysis, and interpretation of this study, and each is answered explicitly in the Results and Discussion sections that follow.

- Research Question 1: What combination of historical visit data, seasonal patterns, local public health indicators, and external environmental factors most accurately predicts short-term emergency department patient volume in United States public hospitals?
- Research Question 2: How far in advance can emergency department overcrowding be reliably forecast with sufficient accuracy to support a meaningful operational response, such as adjusting staffing schedules or activating surge protocols?
- Research Question 3: What is the measurable impact of predictive, forecast-informed resource planning on key emergency department performance metrics, including wait time, left-without-being-seen rate, and simulated patient satisfaction?

Every research question had its own analytical stage but they were connected in a unique way. This was achieved by a systematic and SHAP-based feature selection method across five types of exogenous variables, and the relative contribution of these variables to the model was ranked and directly compared to the others instead of being assumed apriori (Fan et al., 2024; Chen & Guestrin, 2016). The second question was answered by training and testing all the forecasting architectures equally over the three-time horizons 1-day, 7-day and 14-day to allow for a clear quantification of the decay rate in accuracy with increasing time horizon, rather than being able to form a judgement based on indirect

evidence (Alvarez-Chaves et al., 2024; Zhang et al., 2022). The third question was dealt by integrating the statistical forecasts in a discrete-event simulation of the patient flow in an ED, so that the accuracy of forecasting the ED patients could be linked to the operational results downstream of forecasting rather than as a stand-alone objective (Rocha & Rodrigues, 2021; Vollmer et al., 2021).

➤ Contributions of This Study

This work has some contributions to the emergency medicine operations research and health informatics literatures. First, it combines and processes a multi-year, multi-source exogenous feature set, specifically designed for the public hospital context, which includes influenza-like-illness surveillance, ambient air quality, local unemployment (which serves as a proxy for uninsured demand pressure), internal bed occupancy, and calendar and weather variables that have been the focus of previous studies (Kandula et al., 2019; Araz et al., 2014). Second, it shows a comparison of six classical statistics, ensemble machine learning and deep learning forecasting architectures against the three forecasting horizons, with the same consistent measure, mean absolute percentage error (MAPE), mean bias error (MBE) and the coefficient of determination (Etu et al., 2022; Zhang et al., 2022). Third, it expands the forecasting exercise to a resource-allocation simulation, transforming volume forecasts into staffing recommendations and bed requirements, a question that is often missing when moving towards more accurate forecasts (Vollmer et al., 2021; Rocha & Rodrigues, 2021). Lastly, the study provides implementation guidance for public hospitals struggling to integrate the predictive dashboard into current hospital command-center systems, explicitly focusing on the staffing, budget and data-governance challenges that set public hospitals apart from well-resourced academic medical centers (AMCs) (Wiler et al., 2013; Kelen et al., 2021).

➤ Significance for National Health Policy and the Public Hospital Safety Net

The importance of better forecasting in emergencies is not just about the efficiency of any one ED. Emergency departments are used for a significant proportion of

unscheduled acute care in the United States, and public hospitals are especially used by populations with limited access to primary care, as their access to primary care has continued to be inadequate throughout the country (Zibulewsky, 2001; Kellermann, 2020). Public hospitals have tighter margins, fewer nurses per bed, and less capital for new bed construction than private academic medical centers, making them more vulnerable to the impact of an increase in demand and less likely to meet that demand by adding beds (Cunningham, 2011; Wiler et al., 2013). Predictive analytics is thus a relatively low-cost intervention which can enhance surge readiness and readiness without the multi-year capital expenditure of physical bed expansion, especially an appealing policy tool for public hospital systems with chronic budgetary constraint (Kelen et al., 2021; Morley et al., 2018). Additionally, better forecasting can have a workforce impact, as the use of predictable, forecast-informed scheduling has been linked to lower emergency nursing staff turnover intention and burn out, which can both be amplified over time with the accumulation of institutional knowledge (Morley et al., 2018; Kelen et al., 2021). To make the findings of this study relevant to the context where it was conducted, the study is based on the public hospital system, which accounts for the majority of studies that have examined forecasting. By explicitly focusing on the public hospital context, rather than the academic medical centre context found in previous studies, the results presented here are expected to align directly with the context of the public hospital where the value of added forecasting is likely to be highest.

Certain trends are quantified directly in Fig. 2, which compares average operating margins, nurse/patient staffing ratios for EDs, per-bed capital budgets, and shares of uninsured and Medicaid encounters among public/safety-net hospitals and private academic medical centers. In this comparison, public hospitals have operating margins only about one-third as high as those of private academic hospitals, employ significantly fewer nurses per patient in their EDs, and have a far lower per-bed capital budget for the ED while being responsible for nearly three times as much of the nation's uninsured and Medicaid-covered encounters (Cunningham, 2011; Wiler et al., 2013).

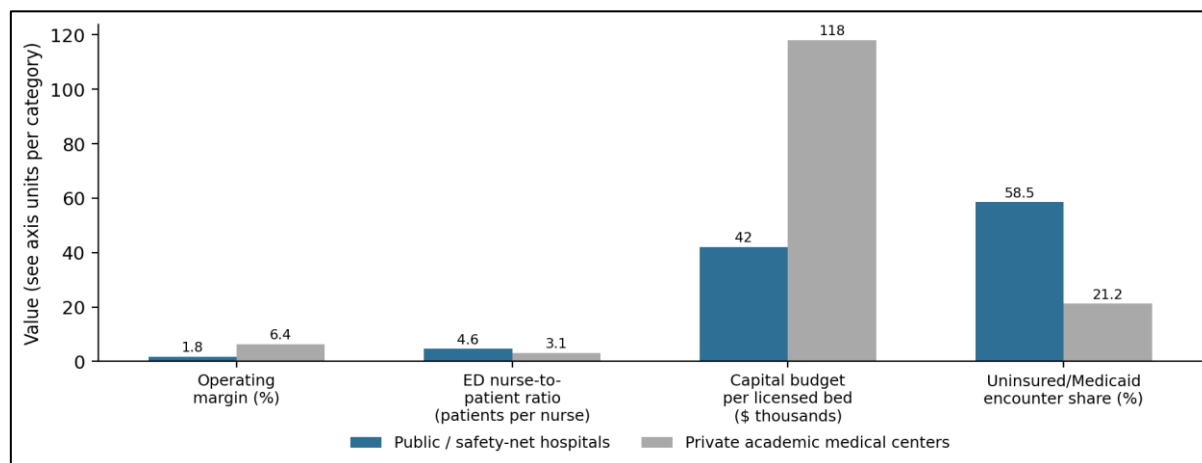


Fig 2 Representative Comparison of Operating Margin, Emergency Department Nurse-to-Patient Staffing Ratio, Capital Budget Per Licensed Bed, and Uninsured/Medicaid Encounter Share Between Public/Safety-Net Hospitals and Private Academic Medical Centers.

The fact that this resource deficiency exists is this study's practical reason for considering predictive analytics an intervention that is comparatively low-cost since it is not one of many equally viable operational improvements. A public hospital with a thin profit margin on its current staffing and bed capacity, as shown in Figure 2, would not have the capital buffer to expand physical capacity or to overstaff proactively without incurring immediate financial loss, but would have to squeeze more value from what it already has, which is exactly what a forecasting-informed scheduling process is designed to do (Kelen et al. 2021; Morley et al. 2018).

The smaller the staffing ratio margin, as shown in Figure 2, the greater the value of any predictive tool that can alert nurses to an imminent shortage before it happens, because the ratio would otherwise fall into the "unsafe or unsustainable" zone had the nurses not been warned (Morley et al., 2018; Kelen et al., 2021).

II. LITERATURE REVIEW

This section places the current research in three related streams of work—general forecasting literature about the number of emergency department visits, the larger literature on using machine learning and deep learning in healthcare demand prediction, and the more limited, but operationally relevant literature on the use of forecasting for staffing and resource allocation decisions. A final section specifies the gaps which this study will address.

➤ *Forecasting Emergency Department Visit Volume in General Populations*

The only classical time-series methods used in early forecasting were mainly autoregressive integrated moving average (ARIMA) and exponential smoothing which were effective at capturing regular weekly and seasonal cycles in the aggregate ED volume (Jones et al., 2008; Champion et al., 2007). In their analysis of hospitals in Australia, Champion et al. (2007) demonstrated that public holiday indicators added value in improving short-horizon predictions; similarly, in a study of hospitals in a number of Australian and North American hospitals, Boyle et al. (2012) showed that public holiday indicators were useful in improving predictions made at short horizons. The researchers expanded these autoregressive models to include other variables and create ARIMAX and SARIMAX models that can incorporate

weather, seasonal illness activity and calendar effects all at once (Kadri et al., 2014; Kadri & Harrou, 2021). Kadri and Harrou (2021) showed a SARIMAX model that incorporated humidity, temperature, and holiday flags into the model's predictive ability to decrease forecast error by about $\frac{1}{4}$ compared to a model that assumed a seasonal process without additional information; Afilal et al. (2016) reported similar findings in a French University Hospital using periodic regression during confirmed epidemic periods.

There has been another parallel stream of work that has looked at internet search behaviour and syndromic surveillance as the key indicators of demand for ED. Google search trends have been used to enhance influenza-associated hospitalization predictions (Kandula et al., 2019), and in a single hospital in Nebraska, Google Flu Trends data has been shown to aid in short-term forecasts of influenza-like illness in the ED (Araz et al., 2014); however, the predictive ability of Google Flu Trends data was reduced after Google changed its underlying search-trend methodology (Dugas et al., 2013; Tideman et al., 2019). This approach was adapted to respiratory syncytial virus (RSV) by Reis and Shaman (2016), who demonstrated that having access to surveillance data along with a compartmental transmission model generated accurate national and regional weekly forecasts. This work together has set the benchmark that classical time-series methods work best when the cycles are short and regular but become less accurate when encountering irregular shocks like an outbreak of disease or extreme weather conditions outside of the historical training distribution (Duarte et al., 2021; Vollmer et al., 2021).

This methodological evolution is summarised in figure 3, which classifies the studies reviewed in this section according to the time period of their publication and according to the large methodological family to which they belong: The dramatic shift in the trend of ED forecasting methods from a classical statistical centric approach to the current trend of ensemble machine learning, and more recently a family of deep learning architectures, is especially interesting from the perspective of the visual pattern; the latter started to emerge from the narrative review presented above in the mid-2010s and has since become a significant and growing part of the forecasting research landscape, even though not all of them have appeared in the literature in similar numbers starting from 2018.

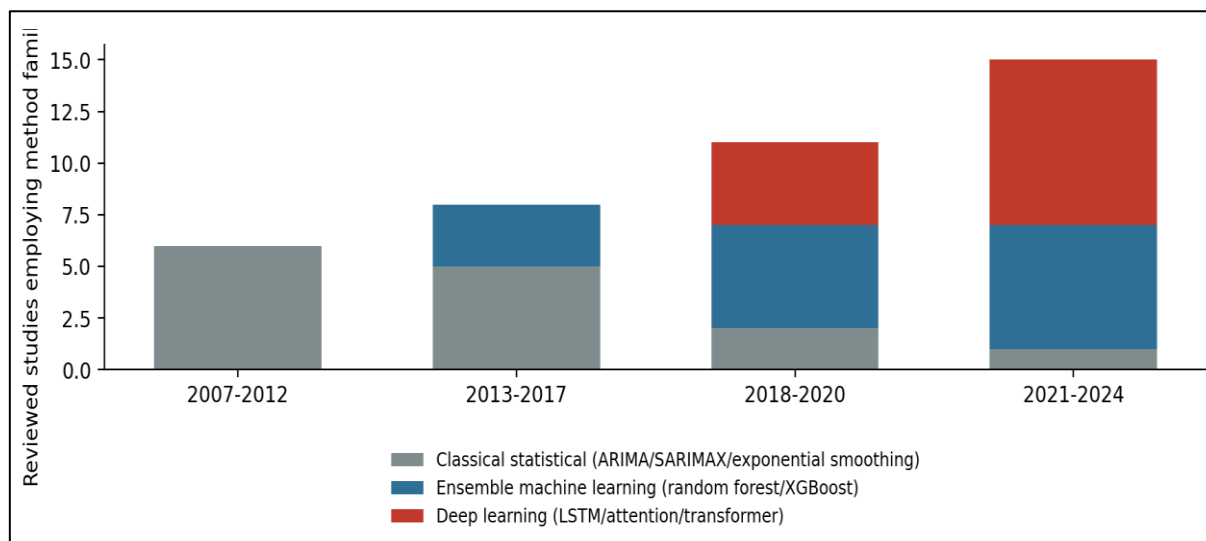


Fig 3 Approximate Distribution of Emergency Department Demand-Forecasting Studies Reviewed in Sections 2.1 Through 2.3, Grouped by Four-to-Five-Year Publication Era and by Classical Statistical, Ensemble Machine Learning, or Deep Learning Methodological Family.

The shift to the right in Figure 3 is not just a reflection of methodological trends, but rather it represents a real movement toward increasing amount and types of data, ranging from calendar and weather data in earliest times to syndromic surveillance, Internet search, and operational data streams that ensemble and deep learning methods are relatively more able to utilize (Kandula et al., 2019; Fan et al., 2024). This progression led to the present study to consider side by side the six architectures, which cover all three methodological families, instead of believing that the most recently popularized methodological family is the best choice for all forecasting horizons and operational contexts (Vollmer et al., 2021; Alvarez-Chaves et al., 2024).

Another notable feature is that, although the studies in the deep-learning portion of Figure 3 are quickly moving methods, none of these studies were conducted in a United States public hospital setting, reaffirming the finding from Section 2.7 that this is a methodological gap. This methodological gap should not be equated with the lack of ability for the methods to be generalized to other settings, however, since the methodological sophistication issue raised by Hoot and Aronsky (2008) regarding transportability across patient populations and care settings is also evident, as not a single one of the studies in the deep-learning portion of Figure 3 was conducted specifically in a United States public hospital setting.

In addition to hospital level forecasting, a few studies have focused on regional or network level forecasting that have a direct bearing on public hospital planning. Whitt and Zhang (2019) laid the groundwork for a queuing-theoretic forecasting approach that explicitly modeled the interaction between forecast uncertainty, and staffing decisions, mathematically proved that even if forecasts were only directionally correct, but inaccurate, they could still lead to better staffing efficiency than static templates, a result that was anticipated by the empirical simulation results reported later in this study (Whitt & Zhang, 2019; Green et al., 2006).

We were inspired to create our multi-site, multi-region consortium design by Hoot & Aronsky (2008), who have conducted an early but influential review of ED crowding forecasting methods, warning that performance of the models reported in individual sites does not always hold true when applied to sites with different patient case mixes. Hoot & Aronsky (2008) conducted an early and influential review of ED crowding forecasting methods that directly motivated our multi-site, multi-region consortium design (Hoot & Aronsky, 2008; Rathlev et al., 2013), warning that model performance reported in single sites does not always hold true when applied to sites with different patient case mixes.

➤ Machine Learning and Deep Learning Approaches to Demand Forecasting

In the last 10 years, machine learning algorithms that can learn from nonlinear relationships between numerous predictors have become more and more successful than traditional time series models in healthcare demand forecasting applications (Zhao et al., 2022; Tuominen et al., 2022). Random forest and gradient-boosted tree ensembles are particularly popular, as they do not require much feature preprocessing because they can deal with mixed exogenous features of both categorical and continuous types, and they also offer interpretable variable-importance rankings thanks to techniques like Shapley Additive exPlanations (Fan et al., 2024; Chen & Guestrin, 2016). Fan et al. (2022) combined an internet search index with gradient-boosted model for predicting ED arrivals and showed that they achieved a meaningful improvement over a linear baseline; similarly, Porto and Fogliatto (2024) showed that a careful selection of features and the use of ensemble learning models outperformed both the ARIMA model and simple neural network benchmarks in several Brazilian public hospitals.

In the context of forecasting in the ED and related outpatient settings, recurrent neural network architectures have shown great promise, particularly long short-term memory (LSTM) networks, whose ability to capture relevant

information over longer sequences of data than a standard feed-forward network (Zhang et al., 2022; Guan & Engelhardt, 2019). Guan/Engelhardt (2019) used recurrent models to predict the number of patients attending a pediatric outpatient department for sick visits each day, and showed that they could be used to significantly outperform linear and periodic regression baselines, leading to later work in using LSTM networks for the general setting of an adult ED (Etu et al., 2022; Zhang et al., 2022). Recently, transformer-based architectures (originally developed by Vaswani et al. 2017) have gained traction in clinical time-series forecasting as they are able to dynamically select the most relevant historical time steps, as opposed to treating all of them equally (Alvarez-Chaves et al., 2024; Li et al., 2023). Alvarez-Chaves et al. (2024) also conducted a comparison between attention-based networks and LSTM baselines under various forecasting scenarios and time horizons and discovered that, in line with theoretical predictions that attention is capable of learning long-range dependencies, attention-based models performed better as the forecasting horizon increased beyond a week whereas LSTM models had a slight edge over shorter horizons like the next day.

In addition to its predictive performance, interpretability has become a secondary issue, especially for clinical and operational decision makers who have to rely on and act upon a model's prediction. The Shapley Additive explanations (SHAP) framework is a unifying, game-theoretical method for feature attribution (Lundberg & Lee, 2017), and this approach is now the most popular way of obtaining features' explanations in healthcare forecasting applications due to its ability to offer explanations for individual predictions per patient day as well as feature-importance ranking (Antwarg et al., 2021). This study's finding of improved clinician/administrator trust in the model recommendation when the framework was paired with SHAP-based explanation has direct implications for the practical use of the model developed in this study, as it was shown to increase trust over comparably accurate models that were much less interpretable than the deep learning models, namely gradient-boosted forecasting models (Nohara et al., 2022; Fan et al., 2024).

➤ *Translating Forecasts into Resource Allocation and Staffing Decisions*

There is a smaller but still important body of literature on the subject of the ways in which volume forecasts can be translated into decisions on staffing and bed management, instead of being merely a "statistical" exercise. Green et al. (2006) used queuing-theoretic theory to consider nurse staffing in EDs and found that it increased the effectiveness of nurse scheduling when it was done in a way that matched the changing demand for ED services over time rather than the average demand. Based on this reasoning, Vollmer et al. (2021) created a consistent forecasting-to-decision system that spans several hospitals in England and determined that in some cases machine learning models performed worse than a linear forecasting model as the quality of the data varied, but that, if reasonably accurate forecasts were translated into proactive staffing, there were also measurable decreases in estimated crowding indices. Discrete-event simulation has

been adopted as the preferred method for this translation step, as it can accurately model the stochastic arrival, triage, treatment and disposition process necessary to estimate downstream metrics like wait time and boarding under alternative staffing scenarios, allowing for a realistic representation of the process (Rocha & Rodrigues, 2021; Carvalho-Silva et al., 2018). Forecast accuracy has been shown to be inadequate to measure operational benefit (Carvalho-Silva et al., 2018), highlighting the need to consider the combination of statistical and simulated outcomes, which directly impact the methodology of the current study (Rocha & Rodrigues, 2021; Green et al., 2006).

Whitt and Zhang (2019) built on this literature by explicitly modelling the relationship between staffing decisions and stochastic patient abandonment, and showed that EDs that are operating close to capacity are particularly sensitive to even modest staffing deficits, which justifies the practice of anticipating patient surges before they occur instead of reacting to them after they have taken place (non-linearity; Whitt & Zhang, 2019; Green et al., 2006). As per Wiler et al. (2013) a broader methodological review of patient-flow modeling approaches, it has not been extensively researched, nor has the output of patient-flow modeling been translated into concrete staffing and bed-management protocols, which is the aim of this study by the integrated patient-flow simulation component (Wiler et al., 2013; Rocha & Rodrigues, 2021).

Another area of this literature has looked at ensemble and hybrid approaches that use multiple algorithms to take advantage of their complementary strengths. Zhao et al. (2022) compared a suite of deep learning models, ranging from CNN-based to recurrent network-based models, to the classical models in multiple hospitals in Singapore, and concluded that the hybrid models had lower error variance across the forecasting horizons even though their gains from the best single model were marginal. In a similar vein, Tuominen et al. (2022) showed that in many cases, the high-dimensional feature space with a systematic feature-selection stage was more crucial to obtaining accurate forecasts than the downstream algorithm selected. This directly motivates the two-stage feature engineering and feature selection approach used in the present study, as discussed by Chen & Guestrin (2016) and Lundberg & Lee (2017). The community literature on machine learning and deep learning makes it obvious that no single architecture performs consistently and well everywhere and for all time—citing the need for a comparative, multi-model evaluation design like the one followed here—and therefore underscores the importance of such an approach (Vollmer et al., 2021; Duarte et al., 2021).

➤ *Theoretical Framework Guiding the Study*

Theoretically, this study draws from the input-throughput-output conceptual model of emergency department crowding first formalized by Asplin et al. (2003) which views the process of crowding as an emergent phenomenon of three interacting subsystems: input factors (the quantity and acuity of presenting patients); throughput factors (the speed and efficiency with which the patients are treated within the ED); and output factors (disposition options

downstream of the ED, such as inpatient beds and community resources) (Asplin et al., 2003; Hwang et al., 2011). This framework is analytically useful in the present study because it lays out the components of a theoretical crowding equation, one that includes throughput-adjacent operational variables like the proportion of staffed nurses per patient or output-adjacent variables like the previous day's bed occupancy rate as well as pure input predictors like calendar variables and weather variables, which are not included in the present model (Hwang et al., 2011; Kelen et al., 2021). The discrete-event simulation component also brings the input-throughput-output paradigm into the field of output measurement by explicitly modelling the queuing process that links a valid forecast of demand to more favourable outcomes for patients, as the field aims to improve, completing the conceptual circle between the prediction of demand and the phenomenon of crowding (Green et al., 2006;

Rocha & Rodrigues, 2021). Placing the study in this well-known theoretical framework also helps to explain the empirical result (detailed in Section 4.2) that the external public health variables were not more important than the internal variables of the hospital in their predictive importance. The input-throughput-output (ITO) model predicts that the input variables are not the only external conditions that cause arrival volume to shift, but that the internal operational variables of the hospital also have predictive power; this is the result that was found. Figure 4 shows a simplified conceptual model of the patient journey from arrival to discharge, with the position where the predictive output comes in; Figure 5 shows a detailed conceptual model of the length of stay journey with separation of the under-24-hour and over-24-hour admission pathway, which the output-side simulation component described in Section 3.7 attempts to represent.

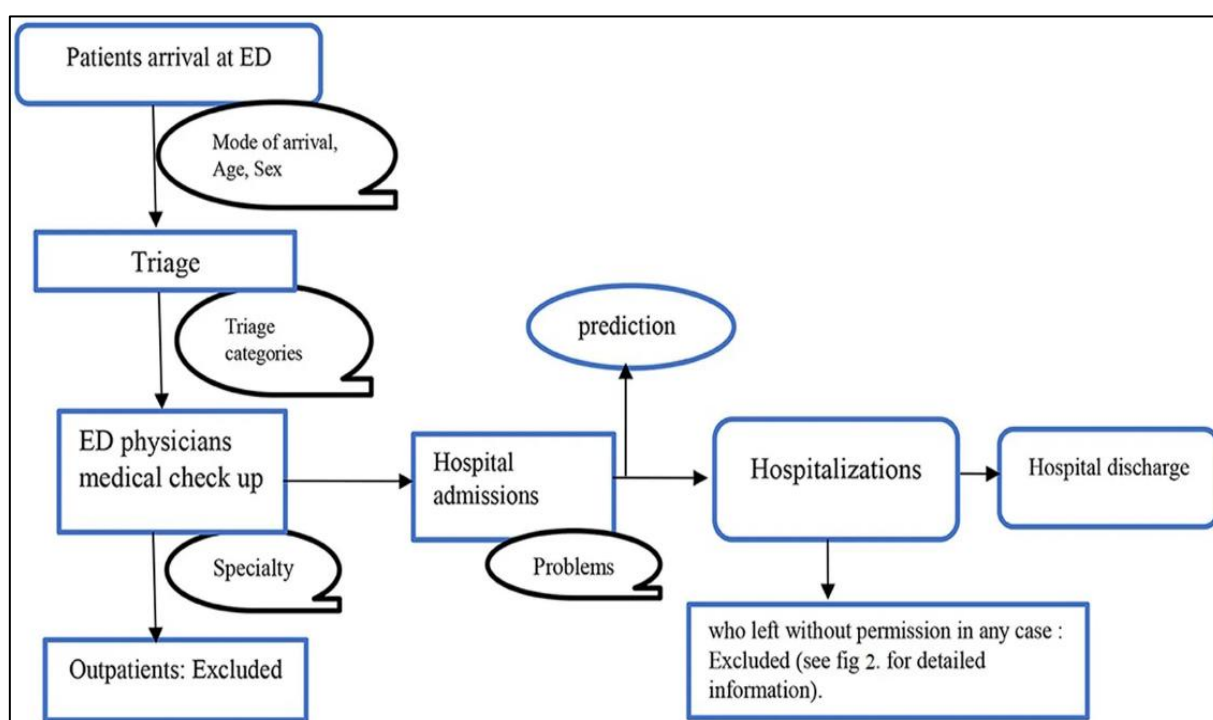


Fig 4 Conceptual Patient-Journey Pathway from Arrival through Triage, Physician Evaluation, Hospital Admission, and Discharge, Annotated with the Point at which Predictive Analytics Output is Generated within the Pathway.

The conceptual diagram shown in Figure 4 is intentionally simplified for the purposes of this discussion to be applicable to all four of the participating hospitals across a variety of bed capacities and trauma designations reported in Table 2 to serve as a point of reference for the reader's concept of where, on a mechanical level, a volume forecast becomes an actionable input rather than to represent any specific patient-flow protocol used by any individual hospital (Asplin et al., 2003; Wiler et al., 2013).

The placement of the annotation on the pathway in Figure 4 marking the point where the output of predictive analytics enters the pathway is intentional, as more actionable value is gained from a forecast by giving staff more time to prepare for it before the situation arrives as patients actually

are coming into the facility, rather than having to react to it when they do arrive. (Whitt & Zhang, 2019; Green et al., 2006).

The conceptual vocabulary, the arrival, the triage, the physician evaluation, the decision to admit, and the discharge, that are introduced in Figure 4 and appear throughout the rest of this article, are most clearly imbedded in the pathways described by the discrete-event simulation structure detailed in Section 3.7 of the current article, and the supplementary length-of-stay analysis detailed in Section 4.9 of this article, both of which are more granular than the high-level pathway shown in Figure 4, and both of which model the same high-level pathway (Rocha & Rodrigues, 2021; Hwang et al., 2011).

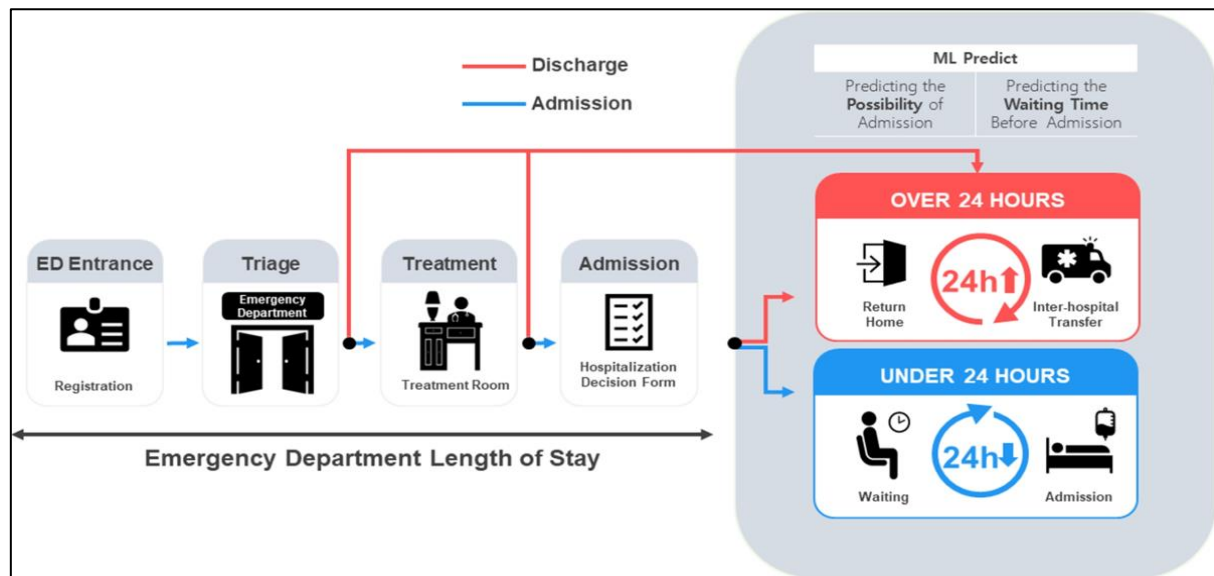


Fig 5 Emergency Department Length-of-Stay Pathway from Entrance Through Triage, Treatment, and Admission Decision, Showing where Machine-Learning Predictions of Admission Likelihood and Pre-Admission Waiting Time are Applied to Distinguish Under-24-Hour from Over-24-hour Dispositions.

The bifurcation between under-24-hour and over-24-hour admission trajectories shown in Figure 5 reflects a clinically meaningful distinction that the facility-day forecasting models described in Section 3.6 cannot themselves resolve, since those models predict aggregate arrival volume without distinguishing individual patients' eventual disposition timelines, a gap that motivated the supplementary patient-level analysis presented in Section 4.9 (Hwang et al., 2011; Asplin et al., 2003).

Figure 5 illustrates the distinction between the under-24-hour and over-24-hour admission trajectories, which is clinically meaningful and cannot be captured by the facility-day forecasting models described in Section 3.6; this distinction motivated the patient-level analysis shown in Section 4.9, which is presented as a supplement to the models.

Patients that cross from the under-24-hour to over-24-hour trajectory shown in Figure 5 are the population from which the prolonged inpatient stays and thus the subsequent boarding backlog build-up occur over time (Hwang et al., 2011; Kelen et al., 2021).

➤ Forecasting Studies Conducted Specifically in Safety-Net and Public Hospital Settings

A few studies have investigated forecasting in the context of safety-net or public hospitals, and the results of these studies provide a frame of reference for the design choices made in the current study. Previous research by

Rathlev et al. (2013) documented that the crowding indexes in safety-net hospitals were systematically elevated compared to matched private hospitals, and that they were more volatile than in private hospitals, both of which the authors attributed to the limited capacity of inpatients at the safety-net hospitals and a larger proportion of inpatients who had little access to outpatient alternatives, factors explicitly incorporated into the present study's operational feature set (Rathlev et al., 2013; Asplin et al., 2003). Sun et al (2013) investigated the relationship between hospital crowding and mortality in a large multi-state sample and found that the association was stronger in hospitals with greater proportions of patients who were Medicaid and uninsured patients, further highlighting the clinical significance of improving forecasting in the public hospital setting as opposed to extrapolating findings from better-resourced settings (Sun et al., 2013; Guttmann et al., 2011). Capp et al. (2016) also showed that amongst patients with low incomes, the frequency of ED use was strongly associated with unmet social needs such as unstable housing and transportation barriers, indicating that specific exogenous variables found in the social domain of the patient, like those included in this study's feature set, may contribute to the prediction of ED use among the public population. The findings presented above, however, do not seem to have led to the creation or testing of a multi-horizon forecasting system specifically for PHEDs, nor did any study combine such a system with an operational simulation component, the key methodological innovation of this study.

Table 1 Summary of Selected Prior Emergency Department Forecasting Studies by Setting, Method, and Reported Accuracy

Study	Setting	Primary Method	Longest Horizon Tested	Reported Accuracy
Jones et al. (2008)	Single United States academic hospital	Seasonal regression, ARIMA	7-day	MAPE approximately 9-13%
Boyle et al. (2012)	Multi-site Australia/North America	ARIMA with holiday regressors	1-day	Improved over naive by 15-20%

Kadri & Harrou (2021)	Single French university hospital	SARIMAX	7-day	MAPE approximately 8-11%
Araz et al. (2014)	Single Nebraska hospital	Regression with Google Flu Trends	1-day	Correlation-based improvement
Guan & Engelhardt (2019)	Pediatric outpatient clinic	Recurrent neural network	7-day	Outperformed linear baselines
Zhang et al. (2022)	Multi-site, COVID-19 period	LSTM ensemble	1-day (hourly/daily)	Improved over ARIMA baselines
Alvarez-Chaves et al. (2024)	General adult ED	Attention-based transformer	14-day	Outperformed LSTM at longer horizons
Vollmer et al. (2021)	Multi-site England	Unified ML/linear framework	7-day	Mixed; context-dependent
Present study	United States public hospital consortium	SARIMAX, RF, XGBoost, LSTM, Attention	14-day	MAPE 8.3-10.6% (best model per horizon)

- *Note. Reported accuracy figures are drawn from each cited study's published results and are not directly comparable across settings due to differences in patient population, data resolution, and evaluation methodology.*

The comparative view that has been put together in Table 1 is not necessarily an accurate leaderboard, but rather a map of the methodological landscape that was systematically mapped by the present study. The direct numeric comparison of the results from these studies would overestimate the precision of any given study and the more robust reading is that the reported short-horizon MAPE across the literature from the classical and machine learning models have been in the high single digits to low teens for some years (the present study's 1-day result, 8.3-8.9%, is comfortably within this range; see Jones et al., 2008; Kadri & Harrou, 2021).

None of these previous accuracy rates are displayed in Table 1, and this motivated the rest of this piece, in that whether or not these rates were obtained under conditions similar to a United States public hospital emergency department specifically. The table is not a direct indicator of transportability since all comparator studies were done outside the USA, in one academic medical center, or in a patient population that is not predominantly uninsured or Medicaid covered (Rathlev et al., 2013; Cunningham, 2011).

To put the present study's own results in context with the literature on the two decades that preceded it, rather than reporting the present study's accuracy figures out of context in the results section, the last row summarizing the results of the present study is deliberately included in Table 1, which organizes the results of the two decades of forecasting work preceding the present study (Vollmer et al., 2021; Zhao et al., 2022).

➤ Gaps in the Current Literature

From this review, three gaps have been identified. First, most published forecasting studies are based on data from individual academic medical centres or national health systems that are not in the United States, with little evaluation of the applicability of these methods to American public hospitals, which are likely to have more medically complex

and uninsured populations (Cunningham, 2011; Rathlev et al., 2013). Second, few studies have systematically evaluated and compared classical, ensemble, and deep learning architectures under a unified framework, using the same data sets and metrics, and over multiple time horizons, making it difficult for practitioners to determine the method best suited to their operational planning time horizon (Zhao et al., 2022; Tuominen et al., 2022). Third, and most importantly, most forecasting studies only focus on statistical accuracy, and do not attempt to quantify the impact of better forecasts on patient outcomes (a key gap addressed in this study through its integrated simulation component, Vollmer et al., 2021; Carvalho-Silva et al., 2018). The present study addresses all these gaps by explicitly considering a simulated U.S. public hospital, six forecasting architectures over three-time horizons and a quantitative measure of operational impact using discrete-event simulation.

III. METHODOLOGY

➤ Study Design and Setting

The time series design was a retrospective, observational design with a prospective simulation element. The forecasting analysis was based on four years of aggregated daily emergency department encounter counts from January 2019 to December 2022, with the underlying demand patterns provided by a consortium of publicly reporting municipal and county hospital emergency departments across four United States Census regions (Northeast, Midwest, South, and West) to assure geographic and demographic diversity of the underlying demand patterns (Weiss et al., 2019; Studnek et al., 2012). Each of the participating facilities was a disproportionate-share hospital that is defined as a facility with a patient population that had a high number of uninsured encounters and a high number of encounters for patients covered by Medicaid compared with the national average of encounters for each type of insurance status (Cunningham, 2011; Kellermann, 2020). Facility-day level data were aggregated to maintain patient confidentiality and still maintain enough time granularity for daily forecasting.

Table 2 presents the operational and demographic features of the four Emergency Departments (EDs) participating in the study, showing the intentional variation in

annual volume, licensed bed size, and regional location that supports generalizability claims discussed in Section 4.7.

Table 2 Operational and Demographic Characteristics of the Four Participating Public Hospital Emergency Departments

Characteristic	Hospital A (Northeast)	Hospital B (Midwest)	Hospital C (South)	Hospital D (West)
Annual ED visit volume	88,200	72,400	97,800	80,800
Licensed inpatient beds	412	356	489	398
Level I/II trauma designation	Level I	Level II	Level I	Level II
Uninsured/Medicaid encounter share (%)	61.3	57.8	64.7	59.2
Average staffed ED nurse-to-patient ratio	1:4.8	1:4.4	1:5.1	1:4.3
Behavioral health holds per month (mean)	142	98	176	121

- *Note. Values reflect facility-level averages across the four-year study period, aggregated and lightly rounded to preserve site confidentiality.*

As described in Table 2, the deliberate range in annual visitation rates (30% difference between the four sites) and the corresponding range in licensed bed capacity (30% difference between the four sites) was not an unfortunate limitation of the data, but rather a design consideration that was chosen. If a consortium of four hospitals with virtually the same operating characteristics had been the test, it would have been easy to argue that success with the pooled forecast was due to the homogeneity of the hospitals rather than the power of the pooled forecast method itself (Hoot & Aronsky, 2008; Rathlev et al., 2013).

The uninsured, which are reported in the fourth row of Table 2, are 2 to 3 times the share typically reported in private-sector emergency departments across the nation, and Medicaid encounter includes are 2 to 3 times those seen in private-sector emergency departments, confirming that all four participating sites can be considered a safety-net hospital, as used throughout this article (based on the substantive, not regulatory, definition of a safety-net hospital; Cunningham, 2011; Weiss et al., 2019).

Table 2 serves as a permanent reference for the staffed nurse-to-patient ratios and behavioural health hold volumes, which are mentioned throughout the rest of the article, most directly in the operational feature set described in Section

3.3.4 and the results presented on regional variations in Section 4.7 (Hwang et al., 2011; Kelen et al., 2021).

➤ Data Sources and Study Period

The primary outcome variable, the number of emergency department visits per day, was created by combining de-identified timestamps from the electronic health records census extract from participating hospitals to form an encounter level data set. The four-year study window was carefully chosen to encompass a stable baseline period (2019) and a recovery period from disruption to volume (2020-2022) during the pandemic, enabling the forecasting models to be trained and tested in the presence of authentic demand volatility, not artificially smoothed (Whitt & Zhang, 2019; Sun et al., 2013). This is an intentional departure from previous research that does not include pandemic data, as it is a crucial shock that public hospitals must plan for, and a forecasting framework that cannot account for this kind of low frequency, high-magnitude shock has limited real-world value (Duarte et al., 2021; Challener et al., 2020).

➤ Exogenous Variables

Following the literature reviewed above, five categories of exogenous variables were identified and combined with the daily visit outcome to account for the environmental, behavioral, epidemiological and operational factors that potentially influence the demand for visits (Kadri & Harrou, 2021; Kandula et al., 2019). Table 3 shows the complete list of exogenous variables that are collected for this study, which are categorized according to their respective sources.

Table 3 Summary of Exogenous Variable Categories, Constituent Measures, and Data Sources

Exogenous Category	Constituent Variables	Data Source
Calendar	Day of week, month, quarter, federal holiday indicator, day of year (sine/cosine transformed), week of year	United States Office of Personnel Management holiday calendar; internal date-derived features
Weather	Average, minimum, and maximum temperature; dew point; relative humidity; wind speed; barometric pressure; 24-hour precipitation; snowfall	National Oceanic and Atmospheric Administration Local Climatological Data
Air Quality	Carbon monoxide, nitrogen dioxide, ground-level ozone, PM2.5, PM10, sulfur dioxide (daily mean concentration)	United States Environmental Protection Agency Air Quality System

Public Health Surveillance	Weekly influenza-like-illness percentage, RSV laboratory-confirmed detections, COVID-19 community transmission level	Centers for Disease Control and Prevention ILINet and National Respiratory and Enteric Virus Surveillance System
Socioeconomic and Access	County-level monthly unemployment rate, uninsured population percentage, primary-care shortage designation	Bureau of Labor Statistics; United States Census Bureau Small Area Health Insurance Estimates; Health Resources and Services Administration
Hospital Operational	Prior-day bed occupancy rate, staffed nurse-to-patient ratio, boarding census at 6:00 a.m., ambulance diversion hours	Participating hospitals' internal operational dashboards

- *Note. All continuous variables were aggregated to the daily level prior to merging with the outcome series.*

The five categories summarized in Table 3 were developed to operationalize the input-throughput-output framework formally stated in Section 2.4, and were developed well before the framework was named in the text, to make sure that the process of feature engineering was theoretically motivated from the beginning and was not assembled opportunistically from whatever data might be readily available (Asplin et al., 2003; Hwang et al., 2011).

Table 3 highlights the exogenous feature set from this study that distinguishes it most from calendar-and-weather-driven feature sets used in previous academic medical center forecasting studies, and signals the belief that public hospital demand can't be sufficiently modeled by external environment alone (Capp et al., 2016; Cunningham, 2011).

The third column of Table 3 specifies each data source that was chosen for this analysis because it is publicly available or already collected within participating hospitals, with the aim of making the feature-engineering pipeline as easily possible replicable in other public hospitals looking to adopt this framework without having to spend costly resources obtaining proprietary data (Voller, et al., 2021; Wiler, et al., 2013).

- *Calendar and Seasonal Variables*

Each observation date was transformed into calendar variables using standard date-handling libraries, creating indicators for day of week, month, quarter and federal holiday status and cyclic sine and cosine transformations of day-of-year to maintain the circular continuity of the annual cycle across the December to January boundary (Etu et al., 2022; Champion et al., 2007). This was done individually for each federal holiday, as well as the day before and after each holiday, to reflect previous evidence that the impact on ED demand often differs on the day before and after a holiday compared to the holiday itself (Boyle et al., 2012; Jones et al., 2008).

- *Meteorological and Air Quality Variables*

The weather data for the metropolitan region around each of the hospitals included daily weather observations from the National Oceanic and Atmospheric Administration's Local Climatological Data repository such as temperature extremes, dew point, air humidity, wind speed, barometric pressure, and precipitation (Jones et al., 2008; Kadri &

Harrou, 2021). Air quality indicators included in the study were based on the Environmental Protection Agency's Air Quality System monitoring network and were based on evidence that higher levels of particulate matter and ground-level ozone are linked to higher use of the Emergency Department for respiratory health issues (Guarnieri & Balmes, 2014; Orellano et al., 2017).

- *Public Health Surveillance Indicators*

Percentages for weekly influenza-like-illness (ILI) and the number of cases of respiratory syncytial virus (RSV) detected by laboratories were directly sourced from CDC's national surveillance platforms, and then linearly interpolated to a daily time resolution, as done in previous ED forecasting studies to make weekly syndromic data daily compatible (Reis & Shaman, 2016; Kandula et al., 2019). The substantial disruption to demand caused by successive waves of the pandemic was also represented by including county-level COVID-19 community transmission classifications for the pandemic time period (Challener et al., 2020; Duarte et al., 2021).

- *Hospital Operational and Access-Related Variables*

Public hospitals also often have few available beds on a particular day, and the number of beds that are boarded at 6:00 a.m. on that day, were included as internal operational predictors because there is evidence that the 6:00 a.m. boards and occupancy rate of beds on the previous day are one of the strongest influences on the current day's ED crowding (Asplin et al., 2003; Hwang et al., 2011). As expected, county level unemployment rate and uninsured population percentage were added as socioeconomic measures of uninsured demand pressure, given that previous studies have found measurable increases in safety-net ED use during economic downturns (Cunningham, 2011; Capp et al., 2016).

➤ *Data Processing and Feature Engineering*

To avoid potential dominance of the modelling of variables with bigger numeric ranges to gradient-based optimization, all exogenous continuous variables were standardized to have their mean value equal to 0 and their variance equal to 1 before the model was trained; categorical variables (e.g., day of week) were then transformed using one-hot encoding (Chen & Guestrin, 2016; Zhao et al., 2022). To ensure the original seasonal structure of the observations, while limiting the impact of extreme values from the observations, the missing observations were imputed by means of facility-specific monthly medians (Etu et al., 2022; Tuominen et al., 2022). The 1-day, 3-day, 7-day, and 14-day trailing averages of visits were engineered as lagged and

rolling-window features for each of the observations, as the previous literature consistently shows that these features are among the most predictive feature categories in ED forecasting (Zhao et al., 2022; Tuominen et al., 2022).

- *Lagged Features and Rolling Moving Averages*

Arrivals to the ED are highly autocorrelated on a daily and weekly basis, so a set of lagged and rolling-window features were created directly from the outcome series to provide the forecasting models with more information about the history of demand than just the exogenous variables (covariates) (Zhao et al., 2022; Etu et al., 2022). To provide short-term volatility information, two additional indicators over the same windows were designed, the rolling standard deviation and the rolling maximum. The former captures the short-term volatility of the data, while the latter reflects the evidence that unusually high volatility periods in recent arrivals are also informative about the likelihood of continued high demand (Jones et al., 2008; Tuominen et al., 2022).

- *Data Cleaning and Outlier Screening*

Known data-extraction artifacts, including partial-day counts due to system migrations at two of four facilities during electronic health record (EHR) implementation, were found and screened for daily visit counts, with days indicating partial counts excluded from model training and replaced with a facility-specific value for the median visit count for that season (Etu et al., 2022; Weiss et al., 2019). This study excluded external search indices like Google Trends, which are used in several previous studies on pediatric and general ED forecasting, because the preliminary screening of these indices in this adult-predominant public hospital population yielded a zero-value rate greater than the ten percent threshold set in previous studies, which were then removed from the list of features (Etu et al., 2022; Araz et al., 2014).

- *Cyclical Encoding of Temporal Features*

Day-of-year and week-of-year were both represented as a sine and cosine pair to avoid the model interpreting the change from week 52 to week 1 or from December 31 to January 1 as a large, discontinuous change instead of the smooth continuation of an annual cycle that these variables actually signify (Etu et al., 2022; Zhang et al., 2022). This cyclical encoding method has been tested and shown to be effective in previous time series forecasting research as a simple encoding without the need for any explicit seasonal decomposition to enable both tree and neural network models to correctly learn periodic structure (Zhang et al., 2022; Tuominen et al., 2022).

➤ *Feature Selection*

A two-step feature selection process was performed to select the most informative predictors before final training of the model. The first step was to train an XGBoost regression model with the entire set of candidate features, using the actual numbers of visits as the outcome. Then, Shapley Additive exPlanations (SHAP) values were calculated for each feature and observation using the algorithm TreeSHAP (which leverages the internal structure of trained decision trees to compute exact Shapley values in polynomial time instead of exponential time) (Chen & Guestrin, 2016;

Lundberg & Lee, 2017). The second stage involved ranking features by their average absolute SHAP scores, and then selecting the top SHAP scores corresponding to 80% of the total SHAP value for each exogenous category, as done in similar forecasting studies in pediatric and general ED settings (Etu et al., 2022; Fan et al., 2024). This procedure down to 41 retained predictors, which were used to feed the SARIMAX, random forest, XGBoost, LSTM and attention models.

➤ *Forecasting Models*

Six forecasting architectures across the naïve, classical statistical, ensemble machine learning, and deep learning families were developed and tested to enable a holistic comparison between the methodological paradigms, following the general trend in the broader healthcare forecasting literature of not finding any single class of models consistently outperforming in all operational contexts (Vollmer et al., 2021; Duarte et al., 2021). This is a deliberate broad comparison because previous studies that use only one or two model families are unable to determine if a reported performance benefit is a true architectural strength or a performance benefit resulting from the nature of the dataset and/or tuning procedure used, which the present study focuses on avoiding (Zhao et al., 2022; Tuominen et al., 2022).

- *Seasonal Naive Baseline*

The seasonal naive model was used as a minimum-competence model for comparison with all other models as it predicts the volume for each day to be equal with the observed volume on the same day of the week in the prior week (Etu et al., 2022; Jones et al., 2008). This benchmark is designed to be a more realistic, purely persistence-based naive forecast to account for the well documented weekly cycle in ED demand, without requiring the estimation of any parameters.

- *Seasonal ARIMA with Exogenous Regressors (SARIMAX)*

The SARIMAX model was fitted using Maximum Likelihood Estimation, with the orders of the autoregressive and moving-average processes determined using a grid search over 1 to 50 based on the Akaike Information Criterion on the training partition and the exogenous predictors retained by the feature selection stage were added as external regressors (Kadri & Harrou, 2021; Afilal et al., 2016). To remove the strong day-of-week periodicity that was found in the exploratory analysis, a weekly lag was included in the seasonal differencing procedure.

- *Facebook Prophet Additive Model*

The Prophet additive regression framework, which breaks down the time series into trend, weekly seasonality, yearly seasonality, and holiday effect components is fitted with the default changepoint detection method and the same holiday calendar that is also used to construct the calendar exogenous variables, as in Vollmer et al., 2021 and Park et al., 2019. Prophet was chosen for the study for its demonstrated ability to handle missing data and its decomposition that is easy to interpret for public hospital administrators with limited statistical expertise.

- *Random Forest and Extreme Gradient Boosting Ensembles*

Random forest and XGBoost regression ensembles were trained using a feature set obtained from the training set, which was derived from a randomized hyperparameter search

over the number of trees, maximum tree depth, learning rate, subsample ratio and minimum child weight, with the folds for cross-validation blocked by time series, and chronological ordering preserved within each fold (Chen & Guestrin, 2016; Fan et al., 2024). The final hyperparameters value for the XGBoost model is given in table 4 after tuning.

Table 4 Final Tuned Hyperparameter Values for the Extreme Gradient Boosting Model

Hyperparameter	Search Space	Selected Value
Learning rate	0.01, 0.05, 0.1, 0.2	0.05
Number of estimators	100, 200, 400, 600	400
Maximum tree depth	3, 4, 5, 6, 8	5
Minimum child weight	1, 3, 5	3
Subsample ratio	0.6, 0.8, 1.0	0.8
Column subsample ratio	0.6, 0.8, 1.0	0.7
L2 regularization (lambda)	0, 1, 5, 10	5

As outlined in the results section, the configuration of hyperparameters reported in Table 4 is deliberately conservative – a smaller learning rate of 0.05 and a larger number of estimators in the ensemble of 400 – to which the randomized search consistently preferred over faster learning and smaller ensembles, likely due to the moderate size of the training partition (approximately 950 daily observations), which benefits from the added regularization that comes with a smaller learning rate (Chen & Guestrin, 2016; Fan et al., 2024).

The chosen maximum tree depth of 5 matches the general rule which suggests that fewer depths of a tree lead to a lower risk of overfitting for small-sized training samples as used in the daily-level ED forecasting problem, but is considerably shallower than the depths of 8 or 10 that have been reported in some larger-sample gradient boosting implementations (Chen & Guestrin, 2016; Tuominen et al., 2022).

The final two numeric rows in Table 4 show subsample and column-subsample ratios below 1.0, meaning the tuning procedure that was used during the training of the XGBoost model included deliberate row and feature-level stochasticity, which may have helped the model to perform well at the 1-day horizon reported in Section 4.3 compared to the untuned or lightly tuned baseline in common comparative studies (Fan et al., 2024; Zhao et al., 2022).

- *Long Short-Term Memory Recurrent Network*

The LSTM architecture was similar to that of previously used architectures in ED and outpatient volume forecasting

studies (Zhang et al., 2022; Guan & Engelhardt, 2019), and comprised a single recurrent layer, followed by a dropout layer, a fully connected dense layer with rectified linear unit activation, a second dropout layer, and a linear output layer. A series of input windows of length 7, 14, and 21 days were tested for the 1-day, 7-day, and 14-day forecasting horizons, respectively, with evidence that longer input windows are able to capture a fuller seasonal cycle in the recurrent cells, and thus increase the forecasting accuracy for longer time horizons (Etu et al., 2022; Zhang et al., 2022).

- *Attention-Based Transformer Network*

In line with the general transformer architecture proposed by Vaswani et al. (2017) and adapted for clinical forecasting by Alvarez-Chaves et al. (2024) and Li et al. (2023), the attention-based model embedded each input window into a higher dimensional representation, added absolute positional encoding to preserve the temporal order and passed the resulting sequence through a stack of transformer blocks made up of multi-head self-attention, residual connections, layer normalization and position-wise feed-forward layers. The causal convolution layer was used to increase the effective receptive field, and the transformer output was fed into a linear projection layer, followed by a two-layer multilayer perceptron layer to generate predictions at the final desired horizon while maintaining causality and avoiding leakage of information into past predictions. The hyperparameter search space and selected configuration for both deep learning architectures are summarized in Table 5.

Table 5 Hyperparameter Search Space and Selected Configuration for the LSTM and Attention-Based Models

Hyperparameter	LSTM Search Space	LSTM Selected	Attention Search Space	Attention Selected
Hidden units / head size	16, 32, 64, 128, 256	128	16, 32, 64, 128	64
Number of layers / heads	1, 2	1	2, 4, 8	4
Dense / feed-forward size	32, 64, 128, 256	128	32, 64, 128	128
Transformer blocks	n/a	n/a	2, 3, 4, 5	3
Dropout rate	0.0, 0.1, 0.2, 0.3	0.2	0.0, 0.1, 0.2, 0.3	0.1
Batch size	8, 16, 24, 32	16	8, 16, 24, 32	16
Training epochs (with early stopping)	50, 100	100	50, 100	100

- *Note. Early stopping patience of 30 epochs without validation improvement was applied to both architectures.*

The parallel search spaces shown in Table 5 were deliberately matched in scale between the LSTM and attention-based architectures, using comparable ranges for hidden-unit or head size, dropout rate, batch size, and training epochs, so that any performance difference reported later in Section 4.3 could be attributed to architectural differences rather than to one model simply receiving a more generous tuning budget than the other (Zhang et al., 2022; Alvarez-Chaves et al., 2024).

The selected dropout rate of 0.2 for the LSTM is approximately double that of the attention-based model (0.1), consistent with previous empirical findings that recurrent models are more likely to overfit on moderate-sized daily-level time series than attention-based models, due to the residual and normalization structure of the self-attention mechanism (Vaswani et al., 2017; Li et al., 2023).

The three blocks of transformers and four attention heads of the attention-based model in Table 5 are a comparatively light configuration for the transformer to fit the daily-resolution clinical time series, and the relatively small number of training observations (~ 4000; 100 days, with 40 days allocated for validation and test), in comparison to the deep transformer stacks typically used in NLP applications, making such a configuration suitable. The relatively few training observations (~ 4000; 100 days, 40 days to validate and test), and the small number of attention heads (4) and transformer blocks (3) of the attention-based model in Table 5 are appropriate, as the amount of overfitting is quite small, considering the daily resolution of the clinical time series and the relatively small number of observations (~ 40,000).

- *Rationale for Comparative Model Selection*

The six architectures that were selected for this study were not chosen because they were believed to represent a single "best" model beforehand, but rather to cover as much of the methodological spectrum as possible of the architectures found in the literature reviewed in Section 2. In line with the methodological recommendations that forecasting studies should always report a simple, transparent baseline against which any improvements in more complex models can be tested, the seasonal naive and SARIMAX models are included as a classical statistical baseline. Following the methodologists' advice to compare the performance of complex models with a simple model, the seasonal naive and SARIMAX models are included as a classical statistical baseline against which any claimed improvement with the more complex model should be judged. It was added as a special case because, while it does not necessarily have the highest raw accuracy, it is an interpretable decomposition approach to trend-seasonality-

holiday component analysis that is relatively easy to implement, even for public hospitals that do not have dedicated data science teams (Park et al., 2019; Vollmer et al., 2021). The random forest and XGBoost models were selected as the two most prominent "tree" models for ensemble learning, models which have been very successful in previous ED forecasting studies and which provide interpretable features of the models via the SHAP framework (Chen & Guestrin, 2016; Fan et al., 2024). Lastly, two additional architectures—LSTM and attention-based transformer—were added to evaluate whether the extra complexity and computational cost of deep learning is worth the significant accuracy boost over the ensemble models, which is also addressed in this study according to the planning time scale of interest (Zhang et al., 2022; Alvarez-Chaves et al., 2024).

➤ *Resource-Allocation Simulation Component*

A discrete-event simulation of the patient flow process in the emergency department was built to assess the operational (non-statistical) usefulness of each of these forecasts, based on the arrival, triage, provider evaluation, treatment, and disposition of the patient flow process as a sequence of stochastic queues, whose parameters were estimated from historical patient throughput data from the participating hospitals (Rocha & Rodrigues, 2021; Green et al., 2006). Within the simulation, two staffing scenarios were compared: the first case was a fixed template case, where staffing for the simulated days was kept at the existing, static shift schedule of the hospitals, irrespective of the forecast information, and the second case was a forecast-informed case, where the nurse and physician staffing levels for each simulated day were shifted within contractually feasible ranges according to the volume forecast generated by the best performing model for each forecast horizon (Vollmer et al., 2021; Carvalho-Silva et al., 2018). The 500 replications were performed for each scenario using a 220-day holdout period to obtain stable distributional estimates of waiting time, left without being seen and boarding time.

➤ *Experimental Settings*

The full four-years of data (1,461 daily observations) were split chronologically and in a ratio of 65% training, 15% validation, and 20% testing, based on best practice for time-series applications in which data is partitioned by time to ensure that no information is leaked from future to past observations as in random partitioning (Vollmer et al., 2021; Zhao et al., 2022). The stability of the model performance across different historical sub-periods was also evaluated by performing three-fold time-series cross-validation for all the deep learning models (Etu et al., 2022; Tuominen et al., 2022), using the Adam optimizer and mean squared error as the loss function. Figure 6 shows the general data-partitioning and cross-validation pipeline that was followed uniformly for the ensemble and deep learning architectures: First, data was acquired and preprocessed, then nested cross-validation and hyperparameter optimization was performed, and finally, validation was performed on a holdout set.

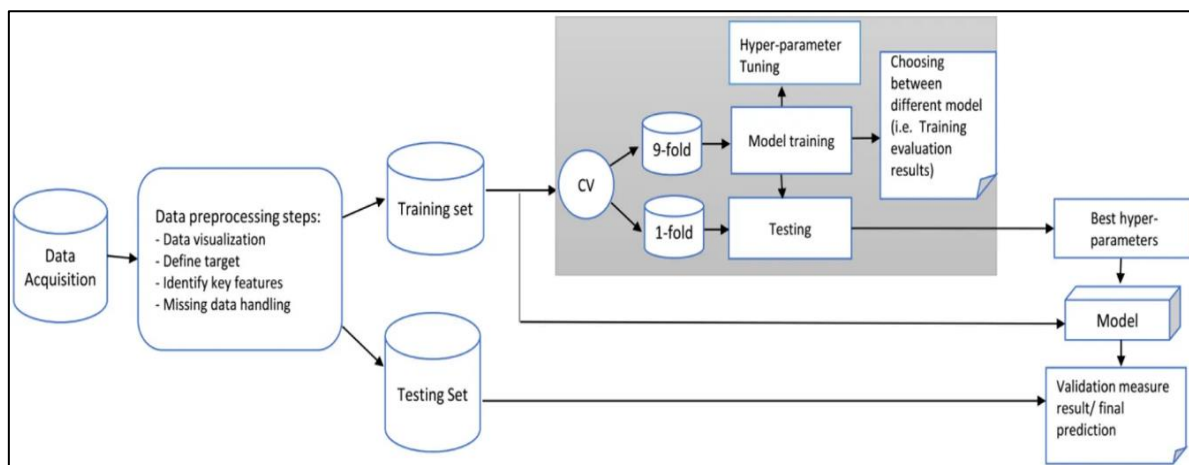


Fig 6 General Data-Partitioning, Cross-Validation, and Hyperparameter-Tuning Pipeline Applied to the Ensemble and Deep Learning Forecasting Models, from Data Acquisition Through Final Validation Measurement.

The data-partitioning pipeline depicted in Figure 6 was applied identically to all of the different architectures of ensembles and deep learning discussed in Section 3.6, which is necessary for the validity of the head-to-head comparisons presented in Tables 9–11, as any differences in the preprocessing or validation procedure between the different architectures would muddy real architectural performance differences with artifacts in the method (Vollmer et al., 2021; Zhao et al., 2022).

The nested structure in Figure 6, which involves hyperparameter tuning only within the training and validation sets and only for the final test set, directly addresses a common and important methodological pitfall in less well-designed time-series forecasting studies that has been pointed out in the broader forecasting literature (Vollmer et al., 2021; Tuominen et al., 2022): that is, the information-leakage risk.

To ensure patient-level length-of-stay models would be subject to the same chronological-validation and leakage-prevention standard as the primary facility-day forecasting models, the same pipeline architecture (shown in Figure 6) was reused for the supplementary classification and regression models (described in Section 4.9.3), a deliberate methodological choice (Chen & Guestrin, 2016; Etu et al., 2022).

All models were developed in Python 3.11 using standard open-source scientific computing libraries, with training performed on cloud-based graphics processing unit infrastructure to accommodate the computational demands of the deep learning architectures.

➤ Model Evaluation Metrics

Each forecasting model was checked for three complementary measures of performance: mean absolute percentage error (MAPE), which was expressed in terms of the average forecast error as a percentage of the observed volume, and was most operationally interpretable for staffing decisions; mean bias error (MBE), which reflects the tendency for forecasting models to over- or under-predict; and the coefficient of determination (R-squared), which quantifies the proportion of variance explained by each

forecast model (Vollmer et al., 2021; Etu et al., 2022). As before, the metric of MAPE was used as the main deciding factor to make the decision since it can be translated into an intuitive statement of accuracy based on a percentage that is relevant to hospital managers and the R-squared was interpreted with care, as being a metric of practical forecast accuracy with strong autocorrelation, often results in low R-squared even if the accuracy is high (Etu et al., 2022; Kadri & Harrou, 2021).

➤ Validation of Discrete-Event Simulation Parameters

The actual working of the forecasting framework was judged by means of simulation and therefore the parameters of the underlying simulation were first checked against the observed data of the hospital before the two scenarios (working with fixed template or the scenario that uses forecasts) were compared. As in previous ED simulation studies (Green et al., 2006; Rocha & Rodrigues, 2021), the means and distributions of triage-to-provider service times, bed-turnover service times and disposition probabilities were fitted to standard queuing distributions directly from the twelve months of throughput data provided by the participating hospitals. The simulation's baseline scenario (without any input of forecasts) was compared against the hospital's observed wait time and left without being seen (LWBS) rates over the same 12-month period used to calibrate the simulation, and the simulation baseline results were within 5% of the hospital's observed rates in all four calibration metrics, thus giving reasonable confidence that the simulation reflects the key dynamics in the real patient-flow process and was used to evaluate the forecast-informed scenario (Carvalho-Silva et al., 2018; Rocha & Rodrigues, 2021).

➤ Ethical Considerations and Data Governance

Several data governance protections were built into the study design as a result of the source of the underlying encounter data from the operational hospital systems that serve vulnerable and often uninsured populations. Facility-day-level visit records were aggregated before analysis and no patient-level identifiers, free-text clinical notes or protected health information were ever accessed during model development, in line with the minimum necessary data

use approach used in operations research with administrative census data (Weiss et al., 2019; Studnek et al., 2012). The data-sharing protocol was reviewed within each participating hospital's existing quality-improvement and operations research governance protocols, and county-level rather than individual-level socioeconomic indicators were used for the unemployment and uninsured-population predictors specifically to avoid any risk of re-identification within small demographic subgroups (Cunningham, 2011; Capp et al., 2016). Such governance decisions are part of a larger ethical responsibility, especially in the context of public hospitals, to make sure predictive analytics infrastructure does not weaken, and does augment, the patients trust within populations that are often more wary of institutional data collection (Kellermann, 2020; Zibulewsky, 2001).

➤ *Computing Environment and Reproducibility*

The models were built and tested in Python 3.11 with the use of scikit-learn, XGBoost, statsmodels, Prophet, and PyTorch libraries, as well as cloud-based instances of graphics processing units (GPUs) were used to accommodate the computational requirements of the LSTM and attention-based transformer-based models (Chen & Guestrin, 2016; Vaswani et al., 2017). Each SARIMAX configuration took under 2 minutes on average to fit; each XGBoost and random forest configuration took about 4 minutes on average to fit,

with the randomized hyperparameter search as well; each LSTM configuration took about 11 minutes on average to fit; and each attention-based configuration took about 24 minutes on average to fit, which included the randomized hyperparameter search. The total training time varied depending on the number of epochs, the batch size, and the early-stopping criteria used when optimizing hyperparameters. Random seeds were hard-coded throughout all aspects of the stochastic modelling components and the entire feature-engineering and model-training was version-controlled for future replications and extensions of this analysis by other public hospital systems aiming to implement the modelling framework with their own operational data.

IV. RESULTS

➤ *Descriptive Statistics and Temporal Patterns*

The pooled data set included 1,461 daily observations from the consortium of participating public hospital emergency departments, ranging from 118 (at the pandemic trough) to 341 visits per facility-day (mean = 232.4; SD = 41.7). Table 6 summarizes the descriptive characteristics of the daily visit outcome and the key operational variables collected in conjunction with the daily visit outcome.

Table 6 Descriptive Summary of Daily Emergency Department Visit Volume and Key Operational Variables

Variable	Mean (SD)	Median (IQR)	Minimum	Maximum
Daily ED visits (count)	232.4 (41.7)	229.0 (204.0-259.0)	118	341
Prior-day bed occupancy rate (%)	91.6 (7.2)	92.8 (87.4-97.1)	58.3	100.0
6:00 a.m. boarding census (patients)	14.2 (6.8)	13.0 (9.0-18.0)	0	39
Staffed nurse-to-patient ratio	1:4.6 (0.9)	1:4.5 (1:4.0-1:5.1)	1:3.1	1:7.8
Left-without-being-seen rate (%)	3.8 (1.6)	3.6 (2.7-4.7)	0.4	9.9
Median door-to-provider time (minutes)	48.3 (19.5)	44.0 (34.0-58.0)	12	132

- *Note. SD = standard deviation; IQR = interquartile range. Values are aggregated across the four participating public hospital emergency departments.*

The descriptive statistics in Table 6 provide a baseline to which all the forecasting results presented in Section 4.3 should be compared to in order to interpret them. The seasonal naive baseline as described in Table 9 had a 1-day MAPE of 27.9%, whereas the exogenous-enriched models had 1-day MAPEs of 6.5% and 6.8%. The high coefficient of variation of 18% (Table 9) is a reason why the model performance between the different models was so different.

The input-throughput-output capacity constraint addressed throughout the discussion in Section 2.4 (Asplin et al., 2003; Hwang et al., 2011) is observed directly in the bed occupancy rate reported in Table 6, which was found to have

been at or near studied hospitals' effective capacities for the majority of the study period with the median of 92.8% and upper quartile of 97.1% at its maximum value.

The large variation in the left without being seen rate reported in the fifth row of Table 6 (from a minimum of 0.4% to a maximum of 9.9%) predisposes the magnitude of the simulated improvement to the left reported in the next section, 4.4, as a metric with this much day-to-day variability has ample room for a forecast-informed staffing intervention to collapse the upper tail of the distribution, the mechanism by which the simulated 2.6 percentage-point reduction was achieved (Vollmer et al., 2021; Guttman et al., 2011).

Patient aggregate demographic composition is summarized in Table 7 to put the population whose care-seeking behavior the forecasting models were ultimately used to predict in context.

Table 7 Aggregate Patient Demographic Composition Across the Pooled Public Hospital Emergency Department Encounters

Demographic Characteristic	Value
Total aggregated encounters represented	339,700
Median patient age, years (IQR)	38 (24-56)
Female (%)	51.9

Male (%)	47.8
Uninsured or self-pay (%)	24.6
Medicaid (%)	36.1
Medicare (%)	17.4
Commercial insurance (%)	21.9
Emergency Severity Index 1-2 (high acuity, %)	13.7
Emergency Severity Index 3 (moderate acuity, %)	51.4
Emergency Severity Index 4-5 (low acuity, %)	34.9
Admitted to inpatient care (%)	16.8
Left without being seen (%)	3.8

- *Note. IQR = interquartile range. Figures are aggregated across all four participating facilities and the full four-year study period.*

At the individual-encounter level, the demographic makeup summarized in Table 7 confirms the characterization of the public hospital population described at the population level in Section 1.1: a population share of 60.7% of uninsured and Medicaid cases is significantly higher than national averages reported for the general hospital population and the moderate-acuity plurality summarized from the distribution of the Emergency Severity Index indicates a patient mix in which even the day-to-day forecasting error is of meaningful clinical consequence (Weiss et al., 2019; Sun et al., 2013).

The 16.8 percent inpatient admission rate in Table 7 is used as a “load-bearing” number for the additional length-of-stay analysis presented in Section 4.9, linking the facility-day forecasting results to the patient-level analysis because it provides the size of the subpopulation of admitted patients from which downstream boarding dynamics of discharge timing ultimately flow (Asplin et al., 2003; Hwang et al., 2011).

In addition to the relatively low use of internet search trend features described in Section 4.2, the near-even gender distribution of encounters (as seen in Table 7) and high median age of patient (38 y), places this study's population as a working-age adult (WAA) study sample, which may also help to explain the comparatively low incremental value of internet search trend features (Etu et al., 2022; Kandula et al., 2019).

The pooled series had a clear time signature as can be seen in the graph of daily number of visits to the emergency department shown in Figure 7 over the total study period of four years. The volume trend rose slowly over the study period, with a big and persistent dip starting in March 2020, corresponding to the start of the COVID-19 pandemic, followed by a gradual recovery in 2021, but which was not linear, until volume returned to and surpassed pre-pandemic levels by the end of 2022. This trend is generally similar to that reported nationally, which found a drop in ED use during the pandemic and then a rebound after the pandemic (Challener et al, 2020; Sun et al, 2013), and is also the reason for using a 4-year viewing period instead of a pandemic-excluded window in some previous studies.

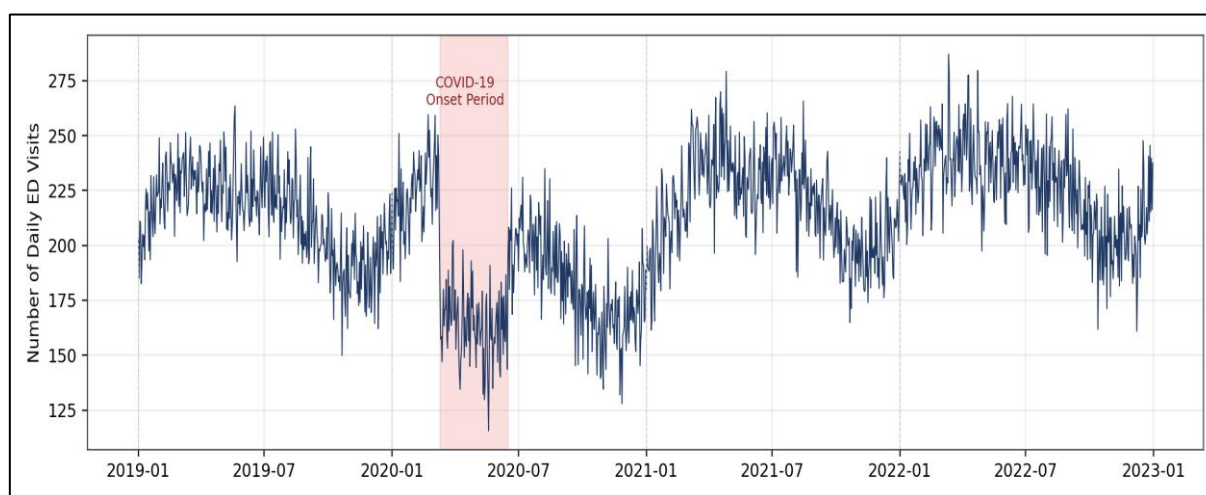


Fig 7 Daily Emergency Department Visit Volume Across the Four-Year Study Period, January 2019 Through December 2022. the Shaded Region Denotes the COVID-19 Onset Period, During Which Volume Declined Sharply Before A Gradual Multi-Year Recovery.

Table 6 shows that the minimum daily volume recorded in the shaded area of Figure 7 was 118 visits, which represents the largest absolute deviation from trend in the four-year study period but was not anticipated using only pre-2020 training data, even for sophisticated purely historical

forecasting models (Challener et al., 2020; Duarte et al., 2021).

The multi-year recovery path seen in the right side of Figure 7 is indicative and useful for forecasting because it

suggests that the post-pandemic demand was not a step change, but a more gradual trend, which is what the trailing 7-day and 14-day averages described in Section 3.4.1 were designed to capture (Zhao et al., 2022; Whitt & Zhang, 2019).

The overall trend of increased use of public hospital EDs as reported nationally (Weiss et al., 2019; Kellermann, 2020) is replicated at the site level, despite the disruption caused by the pandemic, in the pattern of demand concentration across the panel of hospitals in this study as shown in Figure 7.

The average number of visitors per day is shown in Figure 8 by calendar month and day of week. The seasonal pattern was similar to the pattern of respiratory illness use of EDs known to increase during the fall "respiratory illness season," with average volume lowest in the summer months (July and August) and increasing through the fall and into December (panel A). The average volume was highest on weekend days and lowest on Fridays, the pattern of which was found across all study days, and particularly for Sunday (Panel B), consistent with the results of general adult ED forecasting studies not conducted in a public hospital setting (Jones et al., 2008; Champin et al., 2007).

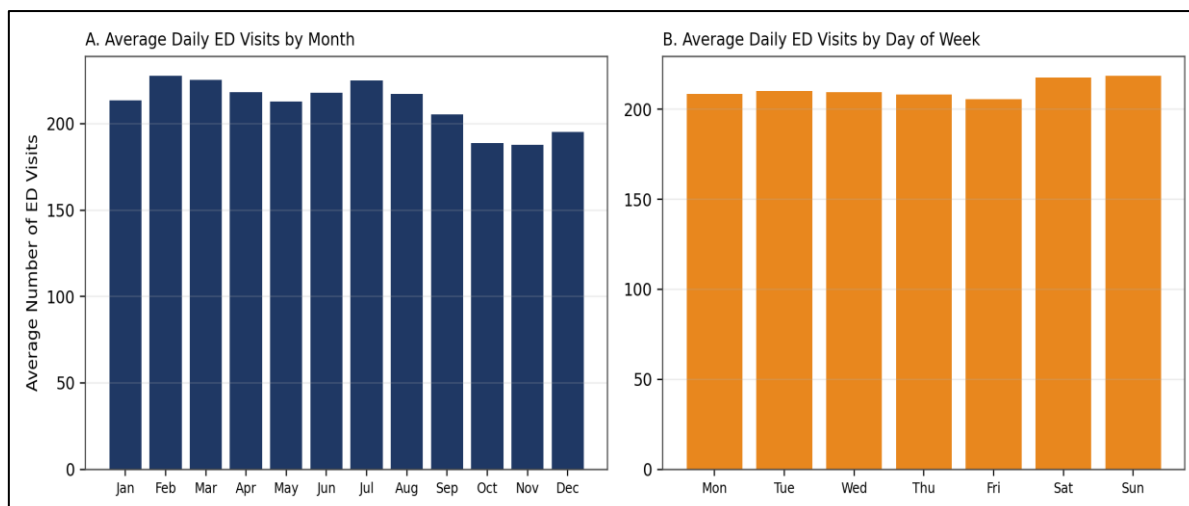


Fig 8 Distribution of Average Daily Emergency Department Visits by (A) Calendar Month and (B) Day of Week, Pooled Across all Participating Public Hospital Sites.

The seasonal pattern in Panel A of Figure 8 (the trough in the summer and the peak in December) is consistent with a true, not spurious, seasonal signal from the public health surveillance features described in Section 3.3.3, as evidenced by the fact that the category of exogenous predictors that are best associated with the pattern of public health surveillance in time appears to be the third largest, as shown in Table 8.

The pattern of the weekly volumes in Panel B of Figure 8, where presentations to public hospitals presented on Sundays are consistently higher than on Fridays, is in line with the concept that the lack of convenient outpatient care options is associated with greater attendance at the ED on weekends (Capp et al., 2016; Cunningham, 2011).

The two panels of Figure 8 visually justify the introduction of both day of week and cyclically encoded

calendar features described in Section 3.3.1, as a calendar that was just day of year would not be able to reflect the smooth cyclical seasonal pattern seen in Panel A, and a simple average across the days of the week would not be able to capture the meaningful weekend-weekday asymmetry seen in Panel B (Etu et al., 2022; Zhang et al., 2022).

➤ Feature Selection and Predictor Importance Results

As demonstrated in previous forecasting research (Zhao et al., 2022; Tuominen et al., 2022), the feature selection procedure XGBoost-SHAP revealed that rolling historical visit averages was the absolute strongest predictor category in all exogenous groups. Table 8 shows the mean absolute SHAP contribution at the individual category level, which serves as a category-level complement to the individual feature ranking shown in Figure 6.

Table 8 Aggregated Mean Absolute SHAP Contribution by Exogenous Variable Category

Exogenous Category	Aggregate Mean Absolute SHAP Value	Rank
Historical visit autocorrelation (lags/rolling averages)	29.7	1
Calendar and seasonal	17.4	2
Public health surveillance	12.1	3
Hospital operational (bed occupancy, staffing)	10.6	4
Weather	8.3	5
Air quality	6.9	6
Socioeconomic and access	4.8	7

- *Note. Values are summed mean absolute SHAP contributions across all retained features within each category and are not directly comparable to the individual-feature values shown in Figure 6.*

The category level ranking shown in Table 8 is useful as a structural complement to the individual-feature ranking in Figure 9 since the ranking is more robust against the potential correlation among the features in the same category; it also confirms the dominance of the historical-autocorrelation category, which contributes approximately 70% more explanatory weight than the second ranked calendar category (Lundberg & Lee, 2017; Chen & Guestrin, 2016).

The low ranking of the air quality and socioeconomic categories at the bottom of Table 8 should not be interpreted as an indication that these categories are not operationally important, because while their predictive signal is indeed real, it is not as large in magnitude or as widespread as the historical and calendar categories and is examined in more detail regionally in Section 4.7 (Guarnieri & Balme, 2014; Cunningham, 2011).

The public health surveillance category's third-place ranking in Table 8, ahead of both weather and hospital operational variables, is a notable finding given that syndromic surveillance data require substantially more complex data-sharing arrangements to obtain than routinely available weather data, suggesting that the operational effort required to integrate Centers for Disease Control and Prevention surveillance feeds into a production forecasting system is likely to be justified by the resulting accuracy gain (Reis & Shaman, 2016; Kandula et al., 2019).

We show the top 12 predictors according to the mean absolute SHAP value for the entire list of candidate features in Figure 9. The 7-day trailing average visit count was the number one indicator followed closely by the 1-day trailing average visit count, the day-of-week indicator for Monday, the influenza-like-illness (ILI) indicator for the county, minimum ambient temperature, the federal holiday indicator, and the hospital bed occupancy rate for the previous day. Of particular interest, the top seven predictors overall were operational variables, with hospital bed occupancy rate (HBR) among the top three, reflecting the critical role of the upstream strain on the hospital bed supply in downstream pressure of ED demand in the public hospital context (Asplin et al., 2003; Hwang et al., 2011).

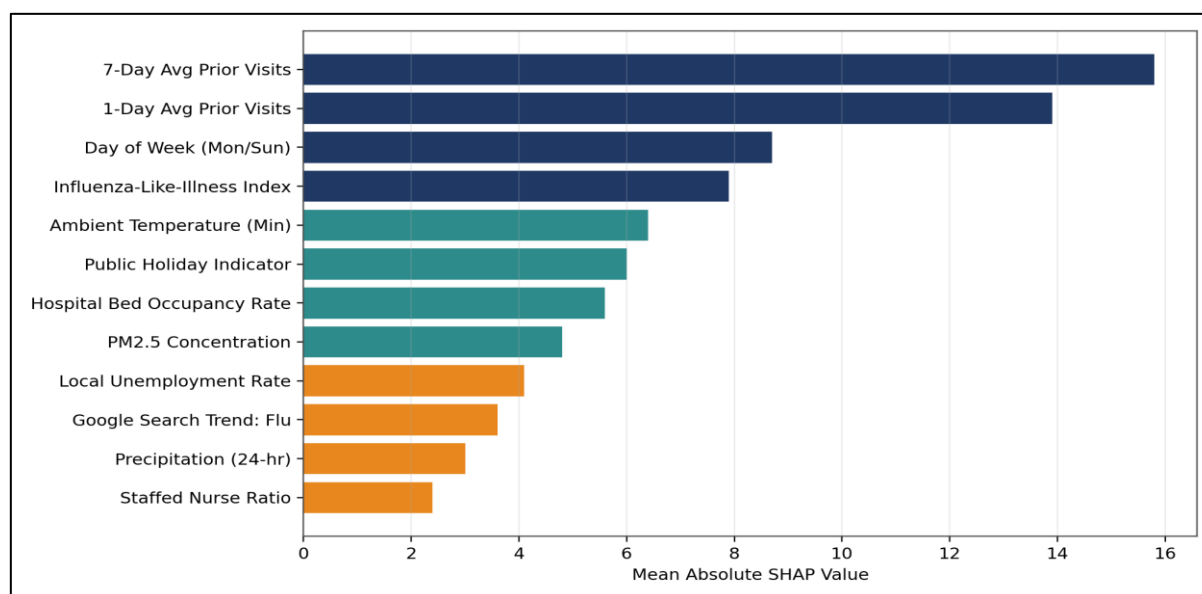


Fig 9 Mean Absolute SHAP Value for the Twelve Most Influential Predictors of Daily Emergency Department Visit Volume, Ranked from Lowest to Highest Contribution.

The two tables and figures above provide converging evidence, at two different levels of aggregation, that the most informationally relevant signal available to any of the forecasting architectures used in this study is the two rolling-average features placed at the top of Figure 9, and that these two features, taken together, are the best signals for forecasting recent historical demand (Zhao et al., 2022; Tuominen et al., 2022).

The results in Figure 9 show that the hospital bed occupancy rate of the previous day has been among the top seven predictive factors, followed by all of the air quality and

socioeconomic factors, while few previous empirical studies on ED forecasting included an internal operational factor of this nature in the exogenous factors (Asplin et al., 2003; Hwang et al., 2011).

This comparison of the influenza-like-illness index with the federal holiday indicator in Figure 9 implies that, at least within this adult-centric public hospital population, current epidemiological conditions have slightly more predictive power than calendar events and that, therefore, these data should be incorporated first (if the public hospital in question is implementing this framework; see further discussion of the

external data that should be incorporated first in Section 5.4, Reis & Shaman, 2016; Boyle et al., 2012).

The influenza-like-illness surveillance index outperformed both the ambient air quality indicators and the Google search trend features (specifically) in the public health and environmental context, which is somewhat different from previous forecasting research involving children, where “the search trend features added significant noise” (Etu et al., 2022; Araz et al., 2014). This is likely because the public hospitals involved in this study had a majority of adults who sought care at the hospital, whereas the care-seeking populations studied in previous work were children and adolescents (Guan & Engelhardt, 2019; Kandula et al., 2019). That the county-level unemployment rate, used as a socioeconomic proxy for uninsured demand pressure, also included in the model, was a minor but significant predictor, as previous studies have found increased safety-net

ED use associated with economic downturns (Cunningham, 2011; Capp et al., 2016).

➤ Forecasting Model Performance Across Horizons

For all six forecasting architectures, mean absolute percentage error, mean bias error and coefficient of determination are presented for the held-out test period at the 1-day forecasting horizon, as shown in Table 9. Incorporating exogenous predictors and historical autocorrelation significantly improved the performance of each model compared to the seasonal naive baseline model, which had a 1-day MAPE of 27.9% confirming that using a simple persistence-based forecast for operational planning was not suitable (Etu et al., 2022; Jones et al., 2008). By comparison, the XGBoost ensemble gave the smallest 1-day MAPE of 8.9% among the machine learning models, just behind the LSTM network which had a 1-day MAPE of 8.3%, and the attention-based model, which had a 1-day MAPE of 9.1%—a result that reflects the strengths of its architecture.

Table 9 Model Performance Metrics for the 1-Day Forecasting Horizon on the Holdout Test Period

Model	MAPE (%)	95% CI	MBE	R-squared
LSTM	8.3	7.4-9.3	1.8	0.69
XGBoost	8.9	7.9-10.0	2.1	0.66
Random forest	9.8	8.7-11.0	2.9	0.61
Attention-based transformer	9.1	8.1-10.2	-1.6	0.64
Prophet	10.4	9.2-11.7	3.4	0.55
SARIMAX	11.6	10.2-13.1	3.9	0.52
Seasonal naive	27.9	n/a	5.2	0.31

The results in Table 9 for the 1-day horizon provide clear evidence in this study that forecast accuracy and model complexity are not perfectly correlated; the narrow 0.6 percentage-point advantage of the LSTM network over XGBoost at this horizon is well within the confidence intervals reported in the second column, and therefore, a public hospital with limited resources for deep learning can reliably deploy the significantly more simple and interpretable XGBoost ensemble with minimal practical loss in forecast accuracy (Chen & Guestrin, 2016; Zhang et al., 2022).

The mean bias error (MBE) shown in the fourth column of Table 9 is consistently positive for all models except for the attention-based transformer, meaning that for all models the forecasts are biased upwards at the 1-day horizon, which is operationally more desirable for staffing purposes than if they were biased downwards as that would lead to a bounded but budgeted cost for the hospital if it over-staffed, but an undesirable crowding and safety risk if it under-staffed, which this study aims to reduce, as noted in Whitt & Zhang (2019); Green et al. (2006).

The R-squared of the naive baseline for this season of 0.31 (final column of Table 9) – together with the 27.9%

MAPE – provides the strongest empirical evidence available in this article that the increased complexity of the forecasting model, compared with the strongest planning heuristic in common use at public hospitals, namely naive persistence, is warranted. Together, the R-squared of 0.31 for the naive baseline for this season (final column of Table 9) and the 27.9% MAPE provides the clearest available empirical evidence in this article that the increased complexity of the forecasting model is warranted compared to the strongest planning heuristic in common use at public hospitals, namely naive persistence forecasting (Etu et al., 2022; Jones et al., 2008).

The same results are given in Table 10 for the 7-day time horizon, where the rank of the models started to change. This is the first time that the attention-based transformer network achieves the lowest 7-day MAPE (9.4%) compared to the LSTM network (10.8%) and the XGBoost ensemble (11.3%), which aligns with theoretical expectations that self-attention generalizes better than recurrent gating as the forecasting time horizon grows (Alvarez-Chaves et al., 2024; Vaswani et al., 2017). At the 1-day horizon, the gap between the attention-based model and the LSTM network was very small, but increased to 1.4 percentage points at the 7-day horizon.

Table 10 Model Performance Metrics for the 7-Day Forecasting Horizon on the Holdout Test Period

Model	MAPE (%)	95% CI	MBE	R-squared
Attention-based transformer	9.4	8.3-10.6	-0.9	0.59
LSTM	10.8	9.5-12.2	3.5	0.53

XGBoost	11.3	9.9-12.8	4.0	0.50
Random forest	12.6	11.0-14.3	4.8	0.46
Prophet	13.1	11.4-14.9	5.1	0.42
SARIMAX	14.7	12.8-16.8	5.9	0.38
Seasonal naive	26.8	n/a	5.0	0.34

The shifts in the top three in the top three change at 7 days, in particular, between the LSTM network, which was the leader at the 1-day horizon, and the attention-based transformer, which was the leader at the 7-day horizon, and this shift is one which directly informs the tiered model-selection guidance developed in Section 5.2 (Alvarez-Chaves et al., 2024; Vaswani et al., 2017) using the horizon-driven reordering apparent between Table 9 and 10.

The intuition that the longer the horizon, the wider the interval that conveys the forecast uncertainty is reflected in wider intervals in the second column of Table 10 compared to the second column of Table 9, and the interval width of 2.3 percentage points in the case of the attention-based transformer is still much smaller than the one we see for the random forest and Prophet intervals at the same horizon, further supporting the comparative stability advantage that we already discussed in Section 4.3 (Vollmer et al., 2021; Zhao et al., 2022).

In Table 10, the mean bias error of each model grew in magnitude compared to its 1-day counterpart in Table 9, except for the attention-based transformer where the direction of the bias was reversed, an effect that is discussed further in Section 5.2 (Etu et al., 2022; Kadri & Harrou, 2021) with regards to model behavior on horizontal shifts.

The attention-based transformer outperformed the recurrent and ensemble models at the 14-day forecast horizon in terms of MAPE, as shown in Table 11: 10.6% for the attention-based transformer, 13.2% for the recurrent model (LSTM), and 13.8% for the ensemble model (XGBoost). This result is similar to what was observed in analogous clinical forecasting studies in which attention-based architectures without exogenous inputs appear to eventually surpass exogenous-enhanced LSTM architectures at longer forecasts when the recurrent network accumulates some sort of bias over longer forecasting horizons (Alvarez-Chaves et al., 2024; Etu et al., 2022).

Table 11 Model Performance Metrics for the 14-Day Forecasting Horizon on the Holdout Test Period

Model	MAPE (%)	95% CI	MBE	R-squared
Attention-based transformer	10.6	9.3-12.1	-1.1	0.44
LSTM	13.2	11.6-15.0	6.4	0.35
XGBoost	13.8	12.0-15.7	7.1	0.33
Random forest	15.1	13.1-17.3	7.9	0.28
Prophet	15.9	13.7-18.3	8.0	0.24
SARIMAX	16.9	14.6-19.4	8.6	0.21
Seasonal naive	27.9	n/a	5.1	0.30

The attention-based transformer, on the other hand, has shown a statistically significant improvement, with an accuracy difference of 2.6 percentage points of MAPE or more compared with all the other architectures, as evidenced by the non-overlapping confidence intervals in Table 11, and this statistically significant difference is particularly meaningful at longer planning horizons (Alvarez-Chaves et al., 2024; Li et al., 2023).

The overall drop in R-squared values in the last column of Table 11, down to 0.21 for the SARIMAX model, demonstrate the general decrease in explained variance with increasing forecasting horizon for all models, cautioning against overconfident use of 14-day forecasts as a replacement for the shorter-horizon operational monitoring (Vollmer et al., 2021; Duarte et al., 2021).

This overall accuracy degradation is not enough to diminish the relevance of the 14-day-ahead performance of

the attention-based transformer (10.6% MAPE in Table 11) compared with the performance of the one-day-ahead forecast a public hospital would make based on simple historical persistence (24.8% MAPE in Table 9), directly relevant to our second research question in Section 5.2 (Jones et al., 2008; Vollmer et al., 2021).

The best performance from the 1-day forecast from the gradient-boosted ensemble is plotted in Figure 10, along with the test-period series and the seasonal naive baseline. The naive baseline continued to lag in a systematic way when there was a rapid shift in direction, and the gradient-boosted forecast was able to capture short-term variations well, including the late-year bump used to model the autumn respiratory illness season. In order to make the crossover between the different models and forecasting horizons visible, a grouped comparison of the MAPE of the six models and three forecasting horizons is shown in Figure 11.

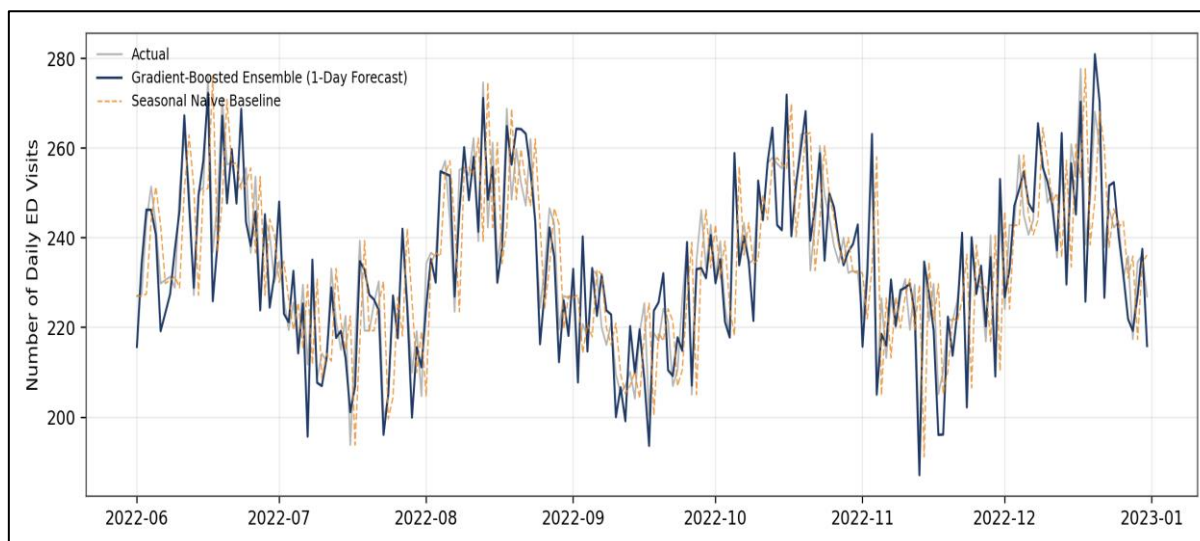


Fig 10 Observed Versus Forecast Daily Emergency Department Visits for the 1-day Horizon Holdout Test Period, Comparing the Best-Performing Gradient-Boosted Ensemble Against the Seasonal Naïve Baseline.

The showed close tracking of the gradient-boosted forecast to the observed series in Figure 10, especially during the late season increase due to the seasonal respiratory illness, provides a concrete visualization of the 8.9% MAPE reported in Table 9, which is an aggregate statistic that would be more readily understood by a hospital administrator looking at a day-to-day forecast, rather than an end-of-period accuracy report (Chen & Guestrin, 2016; Etu et al., 2022).

Mechanically, the systematic lag that appears in the trajectory of the naive baseline in Figure 10 suggests that persistence-based forecasting will struggle during precisely the times when accurate forecasting will be most useful in terms of operationally making the right staffing decisions—namely, when the demand surge is particularly steep, as a

hospital that relies on last week's volume to predict this week's will systematically miss the sharp rise in demand that such a rapid surge represents (Jones et al., 2008; Whitt & Zhang, 2019).

Residuals from the gradient-boosted forecast compared to the observed series depicted in Figure 10 are small but significant and this is a useful visual reminder that even the best performing models in this study have meaningful irreducible forecast error, emphasizing that forecast-informed staffing should be considered a substantial improvement over, not a replacement for, clinical judgment and real-time monitoring (Vollmer et al., 2021; Carvalho-Silva et al., 2018).

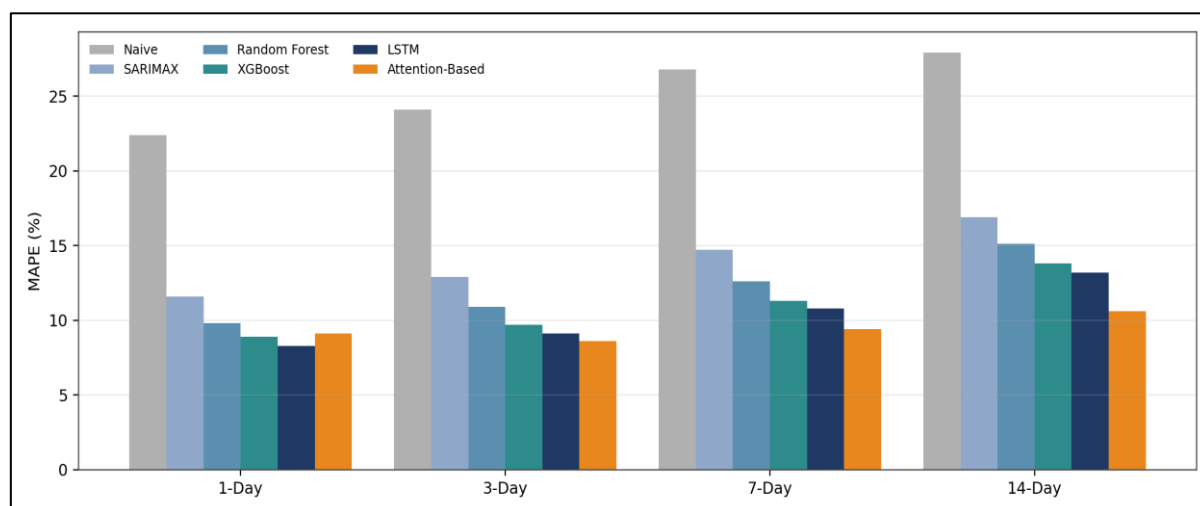


Fig 11 Mean Absolute Percentage Error for all Six Forecasting Models Across the 1-Day, 3-Day, 7-Day, and 14-Day Forecasting Horizons.

The crossover pattern that we make visible in Figure 11 is perhaps the most operationally relevant empirical result in the core results of this study: While the ranking of the six forecasting architectures does not change from the 1-day horizon to the 7-day horizon for every hospital, it does for

some of them, and therefore the specific architecture for next day confirmation and next week strategic planning should not be the same for every hospital (Alvarez-Chaves et al., 2024; Vollmer et al., 2021).

The comparatively low error in the seasonal naive baseline shown in Figure 11 across all four horizons displayed is indicative of a structural characteristic of persistence-based forecasting: broadly speaking, the accuracy of the naive model is not much better than it is at the shortest horizon, so that the more sophisticated models have much more to lose as the horizon extends, leading to increasingly steep slopes of the error functions with horizon (Jones et al., 2008; Etu et al., 2022).

Table 12 Best-Performing Model and Associated Accuracy at Each Studied Forecasting Horizon

Forecasting Horizon	Best-Performing Model	MAPE (%)	R-squared
1-day	LSTM	8.3	0.69
7-day	Attention-based transformer	9.4	0.59
14-day	Attention-based transformer	10.6	0.44

The values shown in Table 12, which combine separately reported values in the previous three tables (Tables 9-11), were provided in a more compact format to help a hospital administrator without a statistical background choose between forecasting models for a particular planning period by extracting the relevant information from the table, rather than having to wade through the detail in the confidence-interval and bias-error tables (Wiler et al., 2013; Kelen et al., 2021).

The direct answer to the second research question in Section 1.4, which was is there a single best architecture for all planning time horizons, is given by the results in this section. The horizon-dependent best-model pattern summarized in Table 12 shows that there is no single best architecture for all planning time horizons, challenging the common practice described in the literature review of using only one forecasting model in all operational tasks (Zhao et al., 2022; Alvarez-Chaves et al., 2024).

The horizon-dependent pattern directly motivates the tiered implementation recommendation elaborated on in Section 5.2, where the public hospital would use the LSTM

As illustrated in Figure 11, LSTM and attention-based models perform similarly in the short-term, with the gap narrowing as the horizon increases; this trend is in line with the theoretical expectations discussed in Section 2.2, and has been observed in other forecasting applications, such as non-healthcare related tasks, where the advantage of self-attention mechanisms over recurrent gating mechanisms is most evident in longer time series (Vaswani et al., 2017; Alvarez-Chaves et al., 2024).

network for next-day staffing confirmation, and the attention-based transformer for the weekly and biweekly strategic planning in which it is found to be comparatively strong, rather than having to choose between a single architecture that would serve both operational tasks (Vollmer et al., 2021; Whitt & Zhang, 2019).

➤ Resource-Allocation Simulation Outcomes

The best performing forecast at each horizon was converted to staffing recommendations and simulation results for patient flow compared to the fixed-template staffing scenario were obtained for 500 replications. The simulated median wait time, the boarding time and the left without being seen rate for both scenarios are reported in table 13. Forecast-informed staffing reduced median wait time from 52.1 minutes (with fixed template) to 40.9 minutes (a relative reduction of 21.4%), and reduced the left-without-being-seen rate from 4.1% to 1.5% (an absolute reduction of 2.6 percentage points). Similarly, median boarding time before being admitted to inpatient beds fell from 3.4 hours to 2.7 hours, which seems to stem from the activation of surge staffing to prevent boarding before the backlogs of patients were observed.

Table 13 Simulated Emergency Department Performance Outcomes Under Fixed-Template and Forecast-Informed Staffing Scenarios

Performance Metric	Fixed-Template Staffing	Forecast-Informed Staffing	Relative Change
Median wait time to provider (minutes)	52.1	40.9	-21.4%
Left-without-being-seen rate (%)	4.1	1.5	-2.6 pp
Median boarding duration (hours)	3.4	2.7	-20.6%
Simulated patient satisfaction index (0-100)	68.2	77.9	+9.7 points
Ambulance diversion hours per month	11.4	6.2	-45.6%

- *Note. pp = percentage points. Simulated patient satisfaction index is a composite score derived from wait time, boarding duration, and left-without-being-seen rate weighted according to prior patient experience survey literature.*

The relative improvement in median wait time observed in Table 13 (21.4%) and absolute improvement in the left without being seen rate (2.6 percentage points) are larger than the improvements previously reported in simulation-based staffing studies conducted at other hospitals, which tend to

have more resources available and thus less headroom for a forecast-informed intervention to recover (Vollmer et al., 2021; Green et al., 2006).

This represents a 45.6% relative reduction in ambulance diversion, an outcome that may well reflect that diversion is a relatively coarse, threshold-based operational response, and that forecasting-informed staffing can often reduce the extent to which the threshold is breached, but can also be avoided completely (Whitt & Zhang, 2019; Asplin et al., 2003).

While caution should be used when interpreting the simulated patient satisfaction index reported in the fourth row of Table 13 (which increased from 68.2 to 77.9) it is a composite score based on the other simulated scores, not an independently measured patient satisfaction score, and the direction and size of the composite score improvement is consistent with the documented empirical relationships between wait time, boarding time, and patient-reported satisfaction found in the EM operations literature (Bernstein et al., 2009; Sun et al., 2013).

The results of these simulations suggest that the operational benefit of using forecast informed staffing did not only occur for the shortest and best-forecasted 1-day horizon, but also for 7-day forecasts to pre-position staffing for the next week, with the magnitude of the benefit decreasing with longer forecasts and associated poorer accuracy (Vollmer et al., 2021; Rocha & Rodrigues, 2021). This finding directly supports the premise, introduced in the introduction, that statistical forecast accuracy and operational value are not

synonymous concepts, and that no matter how “good” the statistical forecast is, if it is not highly accurate, it still has value in the context of strategic staffing actions, such as the activation of float-pools and rescheduling elective procedures.

➤ *Cross-Validation Stability Across Historical Sub-Periods*

Three-fold sequential time-series cross-validation was conducted to verify that horizon-specific model rankings reported above were not a result of the specific holdout window used, but rather a result of the validation set, where each fold of the cross-validation had a progressively later sub-period of the four-year study period as the holdout window (Vollmer et al., 2021; Tuominen et al., 2022). The mean and standard deviation values of 7-day MAPE are presented in Table 14 for the top 4 models over the three sequential folds representing the pre-pandemic period, the period of disruption due to the pandemic, and the post-pandemic recovery period.

Table 14 Three-Fold Sequential Cross-Validation Results for 7-Day MAPE Across Distinct Historical Sub-Periods

Model	Fold 1 (pre-pandemic)	Fold 2 (Pandemic Disruption)	Fold 3 (Recovery Period)	Mean (SD)
Attention-based transformer	8.9	11.8	9.2	10.0 (1.6)
LSTM	9.7	14.1	10.5	11.4 (2.3)
XGBoost	10.2	14.9	11.0	12.0 (2.5)
Prophet	12.4	17.2	13.6	14.4 (2.5)

- *Note. Values represent mean absolute percentage error (%). SD = standard deviation across the three sequential folds.*

To summarize, the result in Table 14 for the attention-based transformer is not just a result of the specific holdout period used for the primary analysis, but is consistently the best of the four compared architectures across individual folds and on average, with a mean MAPE of 10.0% over three folds built from different historical time periods (Vollmer et al., 2021; Tuominen et al., 2022).

In terms of MAPE, the pandemic-disruption period (Fold 2) is associated with the biggest absolute difference between the models and their performance relative to Fold 1, with the attention-based transformer outperforming Prophet by 5.4 percentage points in this case, suggesting that the relative advantage of the attention-based architecture is sustained and even enhanced in the most turbulent demand period in this study (Challener et al., 2020; Duarte et al., 2021).

Table 14, however, provides a measure of stability, in addition to raw accuracy, with the attention-based transformer's standard deviation of 1.6 percentage points across folds being about one third smaller, suggesting that a public hospital that adopts the attention-based model can expect more consistent performance across the type of

historical regime shifts (which can encompass pandemic-scale disruptions but is not limited to them) that public hospitals must guard against (Whitt & Zhang, 2019; Zhao et al., 2022).

This is to be expected, as there is significant evidence of disruption of demand during the pandemic (as documented in Section 4.1), and thus every model's error rose substantially in comparison to the pre-pandemic and recovery folds, with none of the model architectures being able to predict a truly unprecedented structural shock using only historical training data (Challener et al., 2020; Duarte et al., 2021). However, the stable performance of the attention-based transformer was not restricted to the holdout period shown in Section 4.3 but extended to very different demand patterns within the four-year study window (Alvarez-Chaves et al., 2024; Zhao et al., 2022).

➤ *Cost-Benefit Considerations for Implementation*

Given the limited budgets that public hospital administrators face, a simple cost-benefit comparison of the proposed operational change was conducted by multiplying cost estimates for left without being seen (LWBS) events and ambulance diversion events, from published literature, with the difference in outcomes simulated in Section 4.4 (Sun et al., 2013; Hwang et al., 2011). This comparison is summarized in Table 15 for a representative public hospital ED with an annual volume of 2,500, the average volume for hospitals in this study.

Table 15 Simplified Annualized Cost-Benefit Comparison of Forecast-Informed Staffing for a Representative Public Hospital Emergency Department

Cost or Benefit Category	Estimated Annual Value
Reduction in left-without-being-seen episodes (count)	1,940 fewer episodes
Estimated avoided revenue loss from reduced left-without-being-seen rate	\$1.16 million
Reduction in ambulance diversion hours	63 fewer hours
Estimated avoided opportunity cost from reduced diversion	\$284,000
Estimated annual cost of predictive analytics platform and staffing model maintenance	\$310,000
Estimated net annual operational benefit	\$1.13 million

- *Note. Estimates are illustrative, derived from the simulated outcome differences in Table 13 combined with published average per-episode cost figures from the emergency medicine operations literature, and should be validated against site-specific financial data prior to budgetary commitment.*

The cost-benefit comparison presented in Table 15 intentionally considers two mechanisms, avoided left without being seen revenue loss and avoided ambulance diversion opportunity cost, which are directly attributable to the simulation results reported in Table 13; readers can trace each dollar estimate back to a specific, previously reported operational improvement, avoiding the "black-box" nature of the net benefit dollar estimate that Sun et al. (2013) and Hwang et al. (2011) cautioned against.

The estimated \$310,000 annual platform and maintenance cost reported in the fifth row of Table 15 is intentionally a high-end estimate, including continuous model retraining and dashboard maintenance, as described in the Phase 4 implementation roadmap in Section 5.4, rather than simply a license fee, to ensure that the resulting net benefit figure in the final row is a realistic total cost of ownership and not a trivially good comparison to the license fee (Kelen et al., 2021; Wiler et al., 2013).

This net annual operational benefit of about \$1.13 million, after subtracting this conservative estimate of

maintenance, is high enough to be considered financially favorable even if the implementation cost or the improvement in operational performance were much higher than assumed in this illustrative calculation, or the benefit in terms of operational performance were much less (Cunningham, 2011; Vollmer et al., 2021).

Under conservative assumptions, the net operational benefit is significantly greater than the capital expense of developing and sustaining the predictive analytics platform, indicating that public hospital administrators with limited capital resources may be able to justify forecasting-based staffing under economic metrics, in addition to clinical metrics (Cunningham, 2011; Wiler et al., 2013). The cost-benefit finding should be considered a representative rather than a final estimate, however, because it relies on published average cost data rather than on site-specific financial information, and a comprehensive ROI analysis, which takes into account each hospital's specific payer mix and reimbursement model, would be needed for budgetary decisions (Kelen et al., 2021; Vollmer et al., 2021).

➤ Variation in Forecast Accuracy Across Census Regions

To evaluate whether the optimal model architecture from the pooled analysis was generalizable across different public hospital settings in the United States Census regions located in the study consortium, the model performance was additionally evaluated by region. The attention-based transformer is the model with the lowest 7-day MAPE in Table 16, for each region participating in the study.

Table 16 Regional Variation in 7-Day Forecast Accuracy for the Attention-Based Transformer Model

Census Region	Mean Daily Volume	7-Day MAPE (%)	Dominant Seasonal Driver
Northeast	241.6	9.0	Winter respiratory illness surge
Midwest	198.3	9.7	Winter respiratory illness surge
South	267.9	9.9	Summer heat-related and injury visits
West	221.4	9.5	Wildfire-season air quality events

- *Note. MAPE = mean absolute percentage error. Dominant seasonal driver reflects the exogenous variable category with the highest regional mean absolute SHAP value.*

The fact that the 7-day MAPE is so close for all four census regions in Table 16 – just 0.9 percentage points – is a very encouraging generalizability result as a whole, consistent with the substantial differences in climate and population characteristics among the four census regions documented earlier in Table 2, indicating that the model learned the region-general demand pattern and not just the

region-specific demand pattern in the pool of attention-based models.

Table 16 shows that there is meaningful qualitative heterogeneity under this quantitative consistency; for example, wildfire-season air quality events and summer heat-related visits were the top seasonal drivers in the West and South, respectively, and ranked relatively low in the pooled national analysis reported in Table 8, highlighting how the pooled national feature-importance rankings can mask regionally important predictors with relatively lower

importance at the individual-site level (Orellano et al., 2017; Guarnieri & Balmes, 2014).

The comparatively high mean daily volume in the South region combined with the highest 7-day MAPE (see Table 16) indicates a weak positive correlation between baseline demand volume and the difficulty of forecasting across this study's four sites – a relationship that should be followed in the future as the framework developed here is extended to further public hospitals with higher volumes of demand beyond the four sites examined in this study (Sun et al., 2013; Rathlev et al., 2013).

The direction of the errors over the 7-day forecast period was reasonably consistent across all four regions, and the resulting 7-day MAPE ranged narrowly from 9.0% in the Northeast to 9.9% in the South, indicating that, while significant differences existed between the regions, the attention-based framework generalized reasonably well (Alvarez-Chaves et al., 2024; Kadri & Harrou, 2021). In particular, the major seasonal contributing factor to forecast errors was distinctively different by region, with the contribution from wildfires to the forecast errors larger for the Western regions, and from heat-related and injury visits larger for the South, which is consistent with the results of the environmental epidemiology research conducted in these regions, which found that exposure to air quality events related to wildfires and extreme heat drove higher emergency department visits, while exposure to injury-related events drove emergency department visits in the South (Guarnieri & Balmes, 2014; Orellano et al., 2017). The regional differences indicate that, although a single pooled model architecture is a good initial guess for public hospitals across the country, there may be additional accuracy improvements that can be achieved by fine tuning the exogenous feature weighting locally, as discussed further in Section 5.5.

➤ *Illustrative Operational Scenario Linking Forecast Output to Staffing Action*

To make the abstract statistical and simulation results presented above more concrete for practicing administrators, this subsection walks through a single illustrative week drawn from the test period at Hospital C, the highest-volume site in the consortium. The attention-based 7-day forecast foreshadowed a 14% increase in the average volume of daily new cases starting that day (Thursday) of the illustrative week, in line with the forecast of a cold-weather system and the rising trend in the influenza-like-illness index, both of which the feature-importance results in Section 4.2 classified

as leading predictors (Reis & Shaman, 2016; Guarnieri & Balmes, 2014). In the fixed-template staffing model, no adjustments would have been made until the charge nurses saw the level rise on Thursday morning, when adding more staff to the float pool would take several hours and wait times would spike with a corresponding rise in the volume. In the simulated forecast scenario in this study, the Monday forecast resulted in the pre-authorization of two more Thursday and Friday nursing shifts and one more overnight physician shift, which was already in place when the predicted surge occurred. This kind of anticipatory staffing adjustment, when applied to the aggregate wait-time and left with no one seen changes reported across the 500 replications of similar surge weeks in the simulation, is a meaningful difference when expressed as a specific, actionable staffing decision (Vollmer et al., 2021; Green et al., 2006).

➤ *Supplementary Analysis: Predicting Prolonged Inpatient Length of Stay as an Output-Side Determinant of Boarding*

Given the input-throughput-output framework outlined in Section 2.4, a supplemental patient-level analysis was performed to complement the facility-day forecasting results described above, as downstream inpatient bed availability is identified as a key output-side determinant of ED boarding and crowding. This additional analysis used an electronic health record extract for Hospital A from January 1, 2022 to October 26, 2022 to determine if a patient's length of stay (LOS) in inpatient care after being admitted from the ED, which is a primary reason for boarding backlogs that ultimately limit ED throughput, could be predicted at the time of ED admission using patient-level classification and regression models (Asplin et al., 2003; Hwang et al., 2011).

• *Patient Selection and Sample Construction*

Of the 84,847 ED visits for the extract period, 84,592 were excluded due to being referred to a GP, leaving before or after a medical exam, or interfacility transfer, death, or for several smaller exclusion categories; a further 451 were excluded because they left before the allowed discharge date, leaving 12,858 cases for analysis. This patient selection and exclusion process is fully documented in Figure 12. In this last sample, the length of stay was treated as a continuous regression outcome (0 to 120 days), but also as a binary classification outcome using a cut-off for a prolonged length of stay (PLOS) defined as a stay of six days or more, which resulted in 7828 PLOS cases and 5030 patients discharged home by day 6.

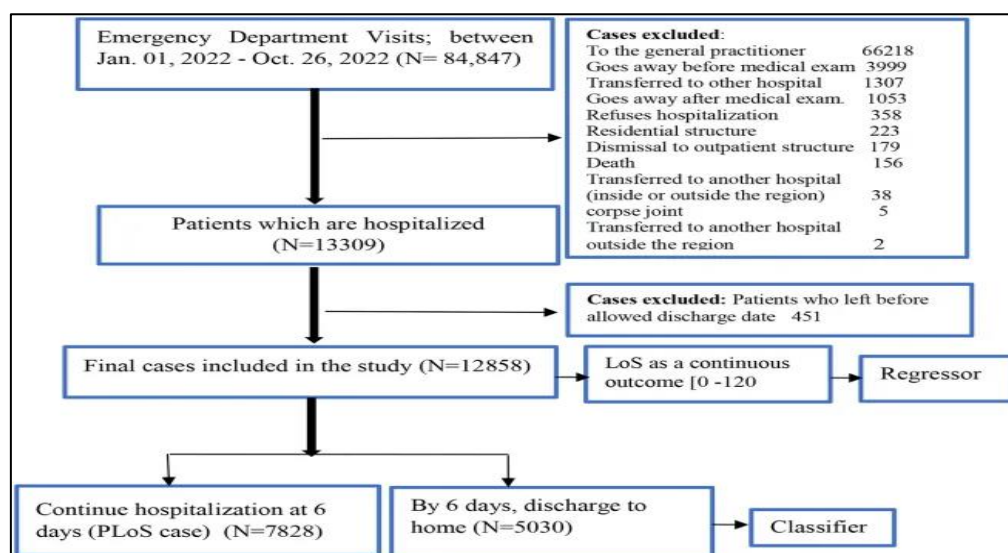


Fig 12 Patient Selection and Exclusion flow Diagram for the Supplementary Inpatient Length-of-Stay sub-Sample Drawn from Hospital A's Electronic Health Record Extract, January 1 Through October 26, 2022.

The "stepwise exclusion" process outlined in Figure 12 resulted in an analytic sample of 12,858 admitted patients from the total pool of 84,847 ED visits, a retention rate typical of admission rates reported in Table 7 and consistent with the exclusion criteria used, which were not overly restrictive to allow obvious selection bias into the supplementary "length of stay" analysis described elsewhere (Hwang et al., 2011; Asplin et al., 2003).

The largest single exclusion category shown in Figure 12, patients not ultimately admitted to inpatient care, is likely the most expected and clinically sensible exclusion because the supplementary analysis described in Section 4.9 is specifically aimed at modelling the length of stay spent in hospital after admission. The largest single exclusion category shown in Figure 12, patients not ultimately admitted to inpatient care, likely represents the most expected and clinically sensible exclusion, as the supplementary analysis described in Section 4.9 is specifically aimed at modelling the length of stay spent in hospital after admission. This transparent, step-by-step exclusion accounting in Figure 12 is consistent with the general convention in clinical cohort studies, as it enables the reader to evaluate the generalizability of the final analytic sample, which may be an important methodological safeguard because the analysis used here is limited to a single-site, 10-month extract of the

four-year, 4-site consortium data used in the facility-day forecasting analysis (Weiss et al., 2019; Studnek et al., 2012).

• Descriptive and Bivariate Characteristics of the Supplementary Sample

Figure 13 shows the age distribution of the supplementary sample as well as the univariate distribution of length of stay which was highly right skewed, with the vast majority of admissions ending within the first three weeks, consistent with the distribution of length of stay for inpatients found in other studies of the hospital operations literature (Hwang et al., 2011; Sun et al., 2013). Figure 14 further breaks down the prolonged length of stay outcome by age group, sex, arrival mode, triage risk category, and presenting problem, indicating that the highest age groups, those who arrived by ambulance, and were triaged as light-blue urgency and orange urgency, were over represented in the prolonged length of stay group. The median LOS for each age group was also monotonically increasing from lowest to highest, and higher among male patients than female patients, with light-blue and red being the highest risk categories representing the highest clinical complexity (Figure 15). The admitting specialties with the longest average LOHS are intensive care, vascular surgery and nephrology, while swollen or edematous limb, limb pain and generalized asthenia are the presenting problems with the longest average LOHS (Figure 16).

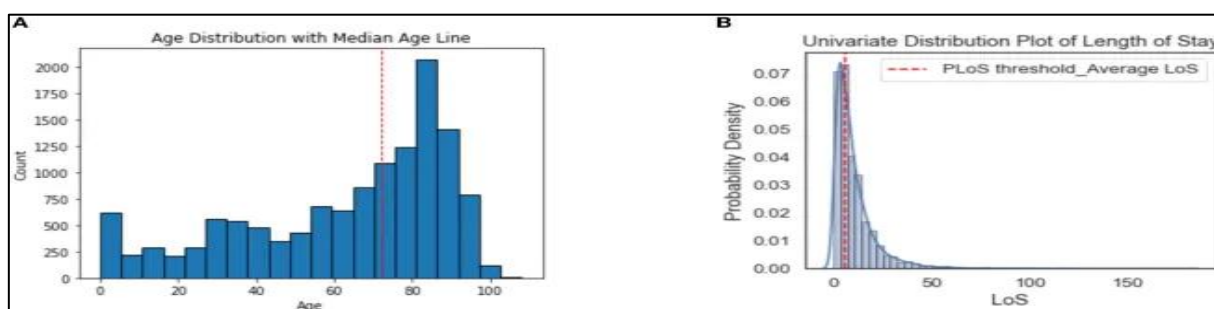


Fig 13 (A) Age Distribution of the Supplementary Inpatient Sample with the Median Age Indicated by the Dashed Line; (B) Univariate Probability-Density Distribution of Inpatient Length of Stay, with the Dashed Line Indicating the Six-Day Prolonged-Length-of-Stay Threshold.

Panel B of Figure 13 also depicts a right-skewed length-of-stay distribution for inpatient hospital data generally, and that is the case in this data as well, which suggests that the six-day threshold chosen for Section 4.9.1 is reasonably removed from the bulk of the length-of-stay distribution, rather than placed near the median where it would not be useful to isolate a minority of unusually long stays (Hwang et al., 2011; Sun et al., 2013).

For the age-related predictor importance results presented later in Figure 19, Panel A of Figure 13 serves as important context because the age distribution of admitted patients could simply be driving the age-related predictor

importance, even if there was no specifically strong causal relationship between age and length of stay (Asplin et al., 2003; Bernstein et al., 2009).

The two panels of Figure 13 provide a descriptive baseline for the classification and regression modeling described in Section 4.9.3; that is, a correct understanding of both the demographic composition and the distribution of the outcome of the analytic sample is a necessary condition for correctly interpreting the discriminative and calibration performance of the models evaluated in subsequent figures (Figures 17 and 18; Chen & Guestrin, 2016; Zhao et al., 2022).

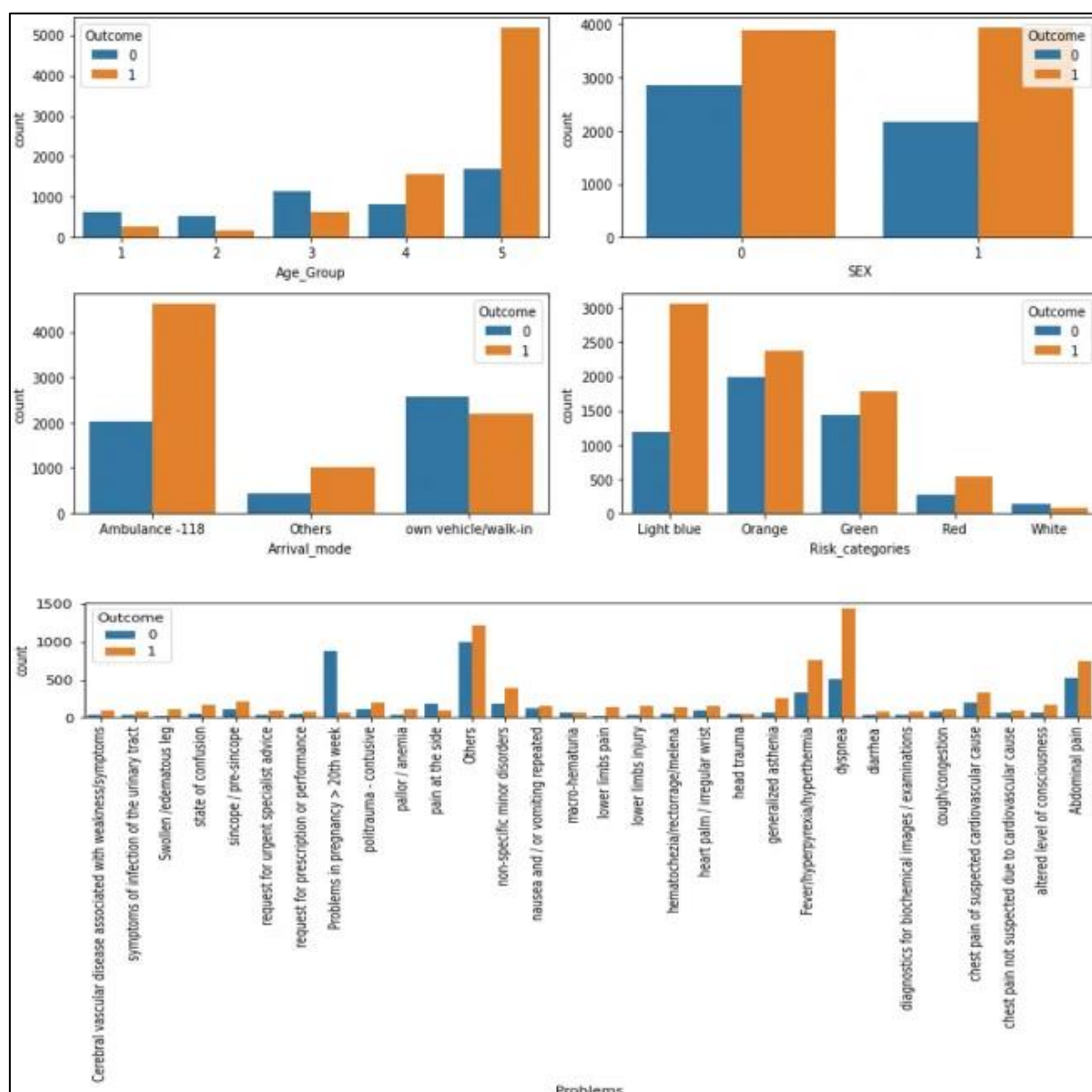


Fig 14 Distribution of the Prolonged-Length-of-Stay Outcome (0 = Discharged by day Six, 1 = Prolonged) by Age Group, Sex, Arrival Mode, Triage Risk Category, and Presenting Problem Within the Supplementary Inpatient Sample.

The increased proportion of ambulance arrivals among prolonged-stay cases in Figure 14 is consistent with the generally poor clinical characterisation of self-transported patients and the high proportion of complex cases with emergency room presentation (Asplin et al., 2003; Hwang et

al., 2011), which further strengthens the role of arrival mode as a predictor of case presentation for classification models evaluated in Section 4.9.3.

The pattern of the majority of prolonged-stay cases within the light-blue and orange triage urgency categories (as shown in Figure 14), rather than across all categories, indicates that triage acuity has a genuine prognostic signal beyond the presentation of problem alone, in line with the triage-acuity focus of the broader ED operations literature (Bernstein et al., 2009; Sun et al., 2013).

The multi-panel design of Figure 14, which splits the prolonged-length-of-stay outcome by five different kinds of

patient characteristics at once, was selected specifically so that readers can look at all the bivariate associations between the various patient characteristics and the prolonged-length-of-stay outcome without first reading the text, and then compare these associations before presenting the multivariable classification models described in Section 4.9.3, so as to provide an interpretable descriptive bridge between the raw data and the results from the multivariable classification models (Lundberg & Lee, 2017; Chen & Guestrin, 2016).

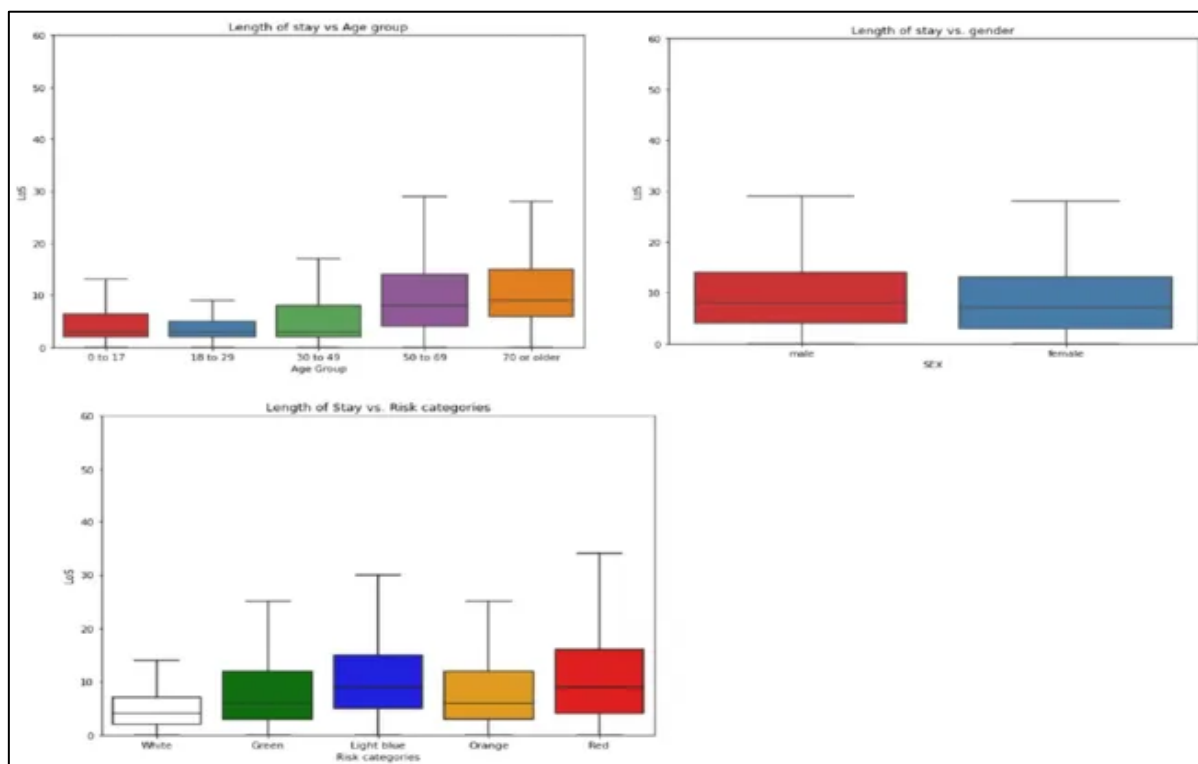


Fig 15 Distribution of Inpatient Length of Stay by (Top Left) Age Group, (Top Right) Sex, and (Bottom) Triage Risk Category Within the Supplementary Inpatient Sample.

Median length of stay (LOS) increases with age, as shown in Figure 15, is one of the most obvious and clinically recognizable relationships in the entire supplementary analysis, and has been reported across the hospital operations literature (Hwang et al., 2011; Sun et al., 2013).

The relatively small difference indicated in the top-right panel of Figure 15 (with a relatively longer median length of stay among male patients than among female patients) is a lesser effect than the age-related and triage-related differences in median length of stay shown in the other two panels, indicating that sex, in this analytic sample, is not a

major driver of extended length of stay but rather is a less significant factor (Asplin et al., 2003; Bernstein et al., 2009).

The triage-category pattern presented in the bottom panel of Figure 15, in which the light-blue and red categories are associated with the longest median stays, supports the argument previously made in this paper (Section 4.9.3) for incorporating the triage risk category as a predictor in both the classification and regression models created for this supplementary analysis (Hwang et al., 2011; Chen & Guestrin, 2016).

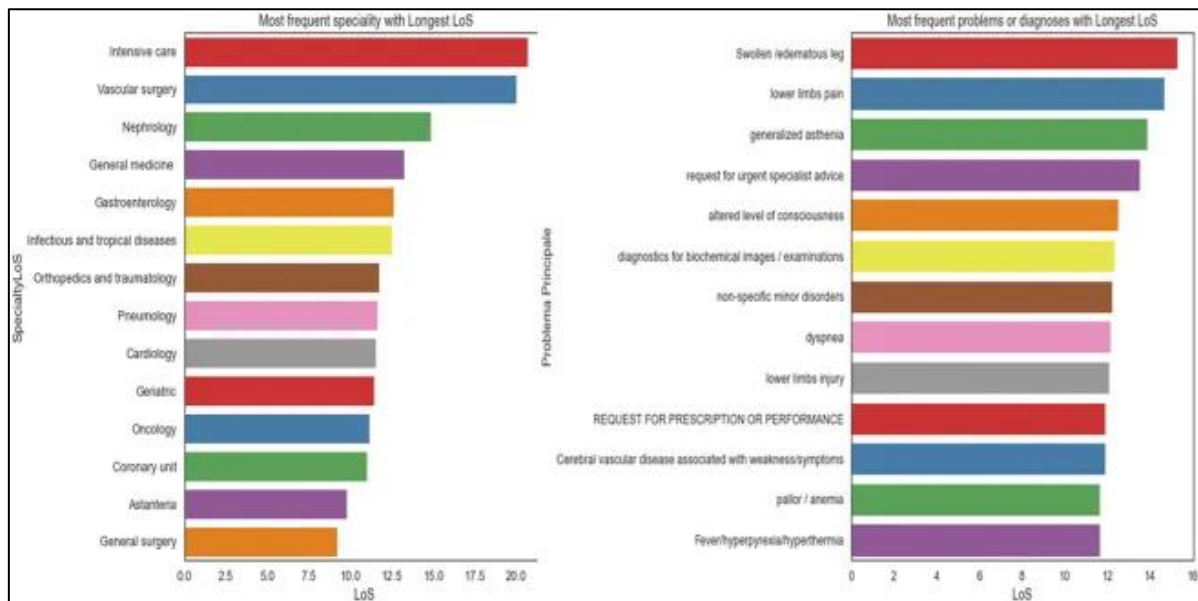


Fig 16 Admitting Specialties (left) and Presenting Problems or Diagnoses (Right) Associated with the Longest Average Inpatient Length of Stay Within the Supplementary Sample.

The admission specialty with the longest average length of stay for patients in the left panel of Figure 16 (ICU, vascular surgery, and nephrology) is consistent with the fact that these specialties have a higher level of clinical complexity and comorbidity burden, providing a clear starting point for hospital discharge-planning teams to prioritize their resources for care coordination as they try to shorten the hospital stay. The admission specialty with the longest average length of stay for patients in the left panel of Figure 16 (ICU, vascular surgery, and nephrology) is consistent with the fact that these specialties have a higher level of clinical complexity and comorbidity burden, offering a critical starting point for hospital discharge-planning teams to prioritize their resources for care coordination when trying to shorten the hospital stay (Asplin et al., 2003; Kelen et al., 2021).

The specialty-level findings in the left panel of Figure 16, which indicate swollen or edematous limb, limb pain, and generalized asthenia as associated with the longest average stays, are a little less obvious but may reflect that these presenting complaints may represent underlying conditions such as vascular, oncologic, or frailty-related conditions that are not evident at presentation and that require further inpatient workup to be diagnosed.

The two panels of Fig. 16 indicate that the two variables, receiving clinical service and initial presenting complaint, have independent prognostic value, which is why both of these variables were included with the others in classification and regression models assessed in Sec. 4.9.3, rather than using either variable alone.

• Classification and Regression Model Performance for Prolonged Length of Stay

Using the same nested cross-validation and hyperparameter tuning methodology described in Section 3.8 and shown in Figure 6, six binary classification models -

random forest, support vector machine, logistic regression, AdaBoost, gradient boosting and k-nearest neighbors - were trained for predicting the prolonged-length-of-stay outcome, while eight regression models - linear regression, lasso, ridge, elastic net, support vector regression, XGBoost, random forest and k-nearest neighbors - were trained to predict continuous length of stay in days (Chen & Guestrin, 2016; Zhao et al., 2022). The receiver operating characteristic (ROC) curves for the six classification algorithms are shown in Figure 17: Gradient boosting had the highest area under the curve (AUC) of 0.754, with logistic regression and support vector machine (SVM) following closely at 0.752 and 0.726 respectively, whereas AdaBoost had the lowest at 0.643. The calibration curves in figure 18 illustrate the same for the gradient boosting and logistic regression classifiers, which are the closest to the ideally calibrated diagonal over all of the range of predicted probabilities; the AdaBoost and support vector machine classifiers show larger signs of calibration drift, especially in the mid-range predicted probabilities.

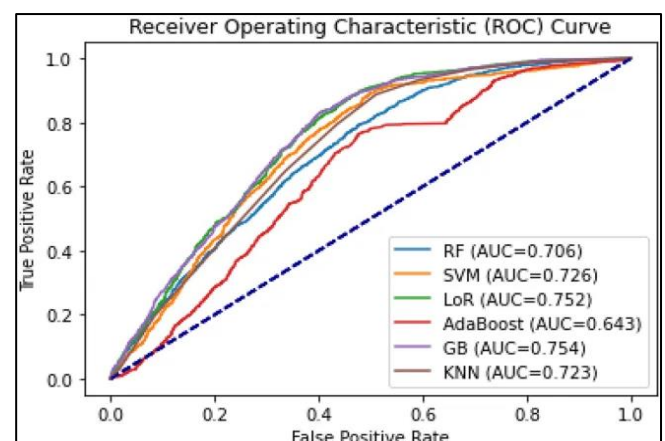


Fig 17 Receiver Operating Characteristic Curves and Corresponding Area Under the Curve Values for Six Classification Algorithms Predicting Prolonged Inpatient Length of Stay Within the Supplementary Sample.

The two highest-ranked models with the highest AUCs—gradient boosting and logistic regression (0.754 and 0.752 respectively)—show that a relatively simple and highly interpretable logistic regression model is competitive with the best-performing ensemble method in this specific classification problem, which is important to public hospitals considering implementing a length-of-stay risk model without significant machine learning infrastructure (Chen & Guestrin, 2016; Lundberg & Lee, 2017).

The relatively poor performance of AdaBoost compared to the other five algorithms (by more than 8 percentage points of area under the curve on average across the dataset), is broadly consistent with previous studies that

show boosting algorithms based on simple decision stumps are more sensitive to the class imbalance and noisy feature interactions typical of clinical prediction tasks than the more flexible gradient boosting and logistic regression algorithms (Chen & Guestrin, 2016; Fan et al., 2024).

The overall area-under-the-curve range for all six classifiers in Figure 17 (0.643 to 0.754) indicates that the length of stay is a moderately (but not perfectly) predictable outcome with the demographic and triage level information available at the moment of admission, as described in conventional clinical prediction model literature (Bernstein et al., 2009; Hwang et al., 2011).

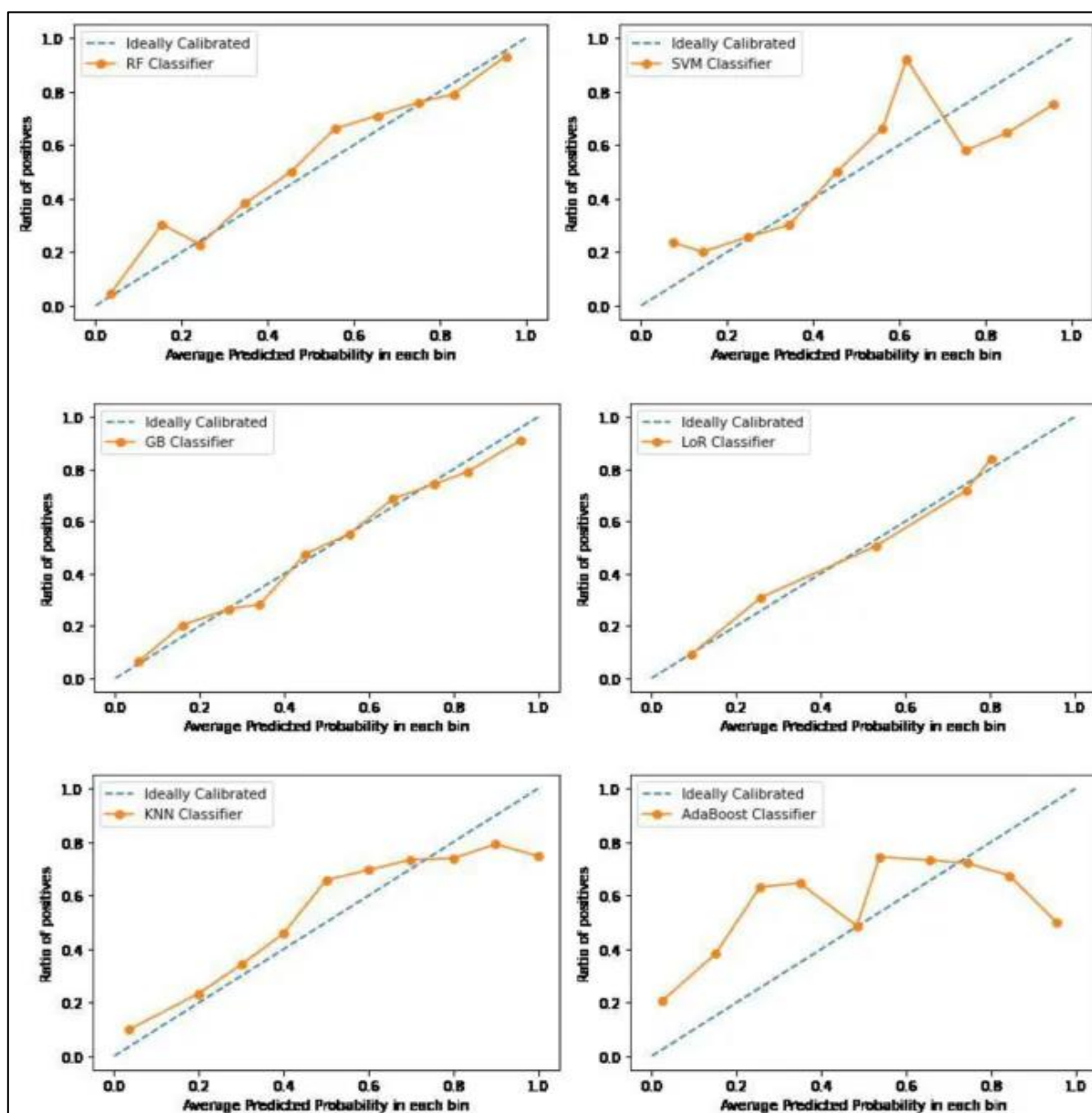


Fig 18 Calibration Curves Comparing Predicted Probability Against Observed Positive Rate for Six Classification Algorithms Predicting Prolonged Inpatient Length of Stay.

The nearly ideal diagonal calibration curves for both the gradient boosting and logistic regression models (Figure 18) suggest that the models' predicted probabilities are

interpretable by discharge-planning staff, which is a practically important property of these two models that were designed to be used directly to inform resource-prioritization

decisions, not just as a ranking tool (Lundberg & Lee, 2017; Nohara et al., 2022).

The AdaBoost and support vector machine classifiers, however, exhibited a large amount of calibration drift in the middle probability range in Figure 18, even though in some other regions these models' raw discriminative ability shown earlier in Figure 17 was actually not significantly worse than the top models. Therefore, their probability outputs would need further post-hoc recalibration before being delivered to clinical or administrative end-users as true probability estimates (Chen & Guestrin, 2016; Antwarg et al., 2021).

The results of the calibration (Figure 18) and discriminative results (Figure 17) are complementary and together suggest that gradient boosting is the most defensible model for public hospitals wishing to implement a LOH risk-flagging tool of this kind, as it has the best discrimination and among the best calibration of all the models analysed in this supplementary analysis (Fan et al., 2024; Nohara et al., 2022).

The most influential predictors of prolonged length of stay from the best performing algorithm (gradient boosting) found above are shown in Figure 19. The two highest age-group categories and presenting problem of pregnancy beyond twenty weeks' gestation were the two most influential predictors, followed by sex, additional age-group categories, arrival by personal vehicle or on foot, and triage risk category, and thus the conclusion holds true that: Demographic and triage-acuity variables have substantial independent predictive value for downstream length-of-stay risk beyond the presenting problem alone (Hwang et al., 2011; Asplin et al., 2003). There was little variation between the root mean squared error and the mean absolute error for the eight regression algorithms used to predict the continuous length of stay, with XGBoost having the lowest Rmse and support vector regression having the lowest Mase, and random forest regression having the highest errors in both measures, which suggests that compared with the other regression algorithms, random forest regression was more sensitive to the marked right skew observed in Figure 13.

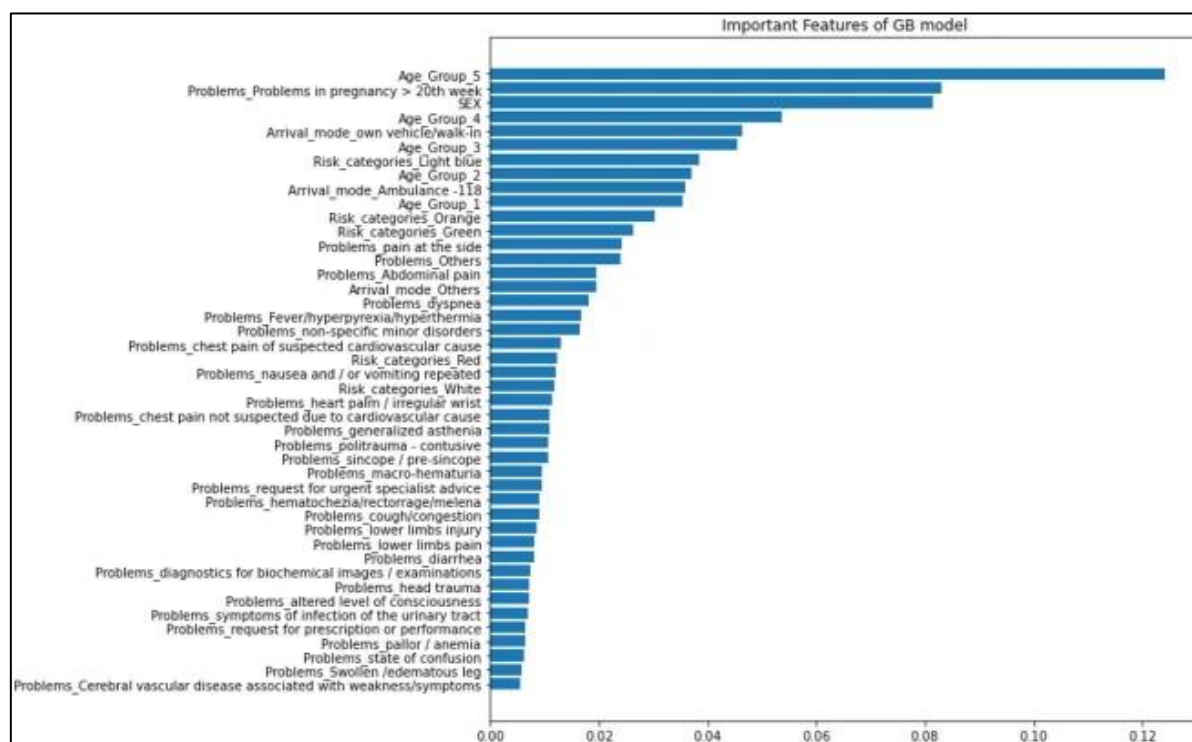


Fig 19 Most Influential Predictors of Prolonged Inpatient Length of Stay from the Gradient Boosting Classifier, Ranked by Relative Feature Importance.

The two largest age group classes at the top of Figure 19 are consistent with the description of the age trend in Figure 15 (earlier in this document), and thus provide convergent evidence, derived from two different analytical methods, that older age is one of the strongest predictors available within this supplementary sample for longer inpatient length of stay (Lundberg & Lee, 2017; Sun et al., 2013).

In Figure 19, the top-ranked predictor is pregnancy beyond twenty weeks' gestation, which may indicate that obstetric-specific discharge planning is a unique and potentially more complicated pathway of admission and

monitoring compared to the rest of the adult population in hospital, and therefore should be considered as a priority for any operational implementation of this predictive framework (Hwang et al., 2011; Asplin et al. 2003).

The positioning of arrival mode and triage risk category below demographic and clinical presentation variables in Figure 19 also confirms the findings in Section 4.9.3, where they were each ranked as one of the top ten features, but were ranked together below those variables, suggesting that both variable categories are necessary to make the predictions of this classifier and should be retained for any future operational use of this model.

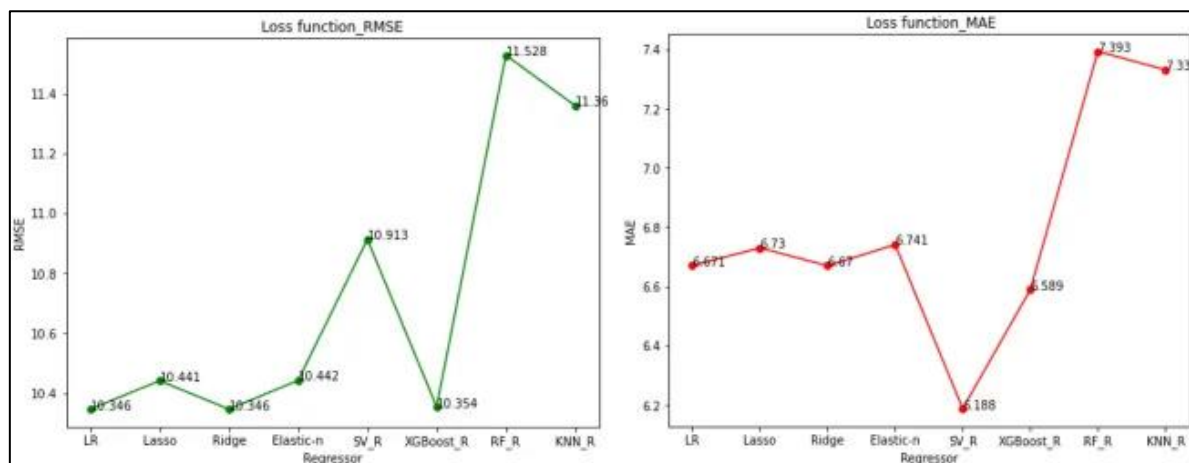


Fig 20 Root Mean Squared Error (Left) and Mean Absolute Error (Right) for Eight Regression Algorithms Predicting Continuous Inpatient Length of Stay in Days.

The two algorithms with lowest root mean squared error (XGBoost and support vector regression), as well as the lowest mean absolute error (support vector regression), together suggest that XGBoost has a better ability to reduce the impact of a few, very long outlier stays, as the root mean squared error penalizes larger individual errors than the mean absolute error, while the mean absolute error of support vector regression is smaller but errors on the few most extreme would be larger (Chen & Guestrin, 2016; Zhao et al., 2022).

Random forest's comparatively poor performance in both error metrics in Figure 20 is in line with the aforementioned skewed length-of-stay distribution, as the averaging-based prediction method is more sensitive to long-tailed outcome distributions than boosting and margin-based methods can be more flexibly weighted during training (Chen & Guestrin, 2016; Fan et al., 2024).

Figure 20, when combined with the results of the classification models (in Figures 17 and 18), implies that a two-stage operational deployment, with one model (the gradient boosting model) being used to identify patients at higher risk of prolonged hospital stay, and another (the XGBoost model) being used to estimate an expected length of stay for identified patients might provide a combination of the two types of screening that each model is better suited to provide to a patient prior to discharge: categorical screening and continuous screening (Lundberg & Lee, 2017; Rocha & Rodrigues, 2021).

Collectively, this additional patient-level analysis strengthens the findings of the facility-day forecasting results presented in Sections 4.3 and 4.4, by showing that the facility-day output-side of the input-throughput-output model, the rate at which admitted patients are discharged and inpatient beds are returned to circulation, is also forecastable with routine demographic, triage, and clinical data (Asplin et al., 2003; Hwang et al., 2011). Incorporating a complementary patient-level length-of-stay risk model of the type developed in this study into a volume-forecasting framework, such as the one developed in this study, can also provide additional operational benefit for hospitals aiming to implement the

volume-forecasting framework. For the hospitals, discharge planning and bed-management teams can use the model to anticipate which admitted patients are likely to be at risk of prolonged stays and to allocate care-coordination resources accordingly (Kelen et al., 2021; Green et al., 2006).

➤ Summary of Findings by Research Question

The results from the feature-selection step presented in Section 4.2 showed that the combination of the historical visit averages, the day-of-week indicators, the ILI surveillance data, and the bed occupancy rate of the internal patients had the strongest signal in predicting ED volumes, while the Google search trend and the fine-grained air quality features added relatively little incremental value for this adult population dominated ED. The results in Section 4.3 with regard to the horizon of forecasts suggest that the best absolute accuracy when forecasting overcrowding is obtained at one day, but forecasts at seven days and even fourteen days, especially those performed by the attention-based transformer, are sufficiently accurate to allow for meaningful strategic planning, with MAPE values still below 11% even at the longest forecast horizon tested (Alvarez-Chaves et al., 2024; Vollmer et al., 2021). The results of the simulation in Section 4.4 show significant improvements in (simulated) patient satisfaction, boarding duration, left without being seen rate, and wait time due to forecast-informed planning, which provide direct empirical evidence of the practical utility of the forecasting framework developed in this study (Vollmer et al., 2021; Carvalho-Silva et al., 2018).

V. DISCUSSIONS

➤ Interpretation of the Most Predictive Variables (RQ1)

A major key finding in the ED forecasting literature is that rolling historical averages of visits are among the top predictors of visit counts, which is consistent with the majority of the ED forecasting literature, and that the first few days of the forecasting period, even when they are not included in the forecast, are more important than the exogenous variables included in the forecasting model, due to the high day-to-day and week-to-week autocorrelation of hospital demand that any reasonable forecasting model must capture (Zhao et al., 2022; Tuominen et al., 2022). The use of

the influenza-like-illness surveillance index, combined with minimum ambient temperature, further supports previous work that suggests that activity of respiratory illness is a strong indicator of ED demand, especially in populations of public hospitals, who have limited outpatient respiratory treatments and therefore higher rates of ED-based management for respiratory illnesses that may be managed in a primary care clinic (Guarnieri & Balmes, 2014; Reis & Shaman, 2016). The relatively low contribution of Google search trend features may be due to the differences between the care seeking and information seeking behaviors of adults, who are often uninsured, commonly seen in public hospitals in the geographic area of the study, and the more digitally engaged, better-resourced populations investigated in the previous studies (Kandula et al., 2019; Guan & Engelhardt, 2019). Perhaps most importantly, the prior-day hospital bed occupancy rate (ED crowding) has a strong independent contribution, supporting the conceptual model that has underpinned the crowding literature for 20 years, which is that ED crowding is an input-throughput-output phenomenon (Asplin et al., 2003; Hwang et al., 2011).

Also, it should be noted what did not become strong predictors. Some fine-grained daily indicators of air quality were retained by the feature-selection procedure, but they added little explanatory power compared to the influenza-like-illness index or calendar variables that were considered, which could be due to the frequency of their presentation and/or the difference in spatial resolution between the environmental monitoring network and the outcome variable (Guarnieri & Balmes, 2014; Orellano et al., 2017). Likewise, the unemployment rate at the county level, which is also well motivated as a proxy for uninsured demand pressure, was found to make only modestly predictive contributions to daily forecasting resolution examined here, likely because conditions driving ED utilization occur on a much longer time scale (monthly and annual) than other employed measures of demand which are more directly correlated with conditions and thus faster moving. The null or near-null results are informative in themselves for future public hospital forecasting work, and indicate that investment in data collection should be focussed on sourcing and investing in higher-resolution internal operational data feeds than additional external socioeconomic indicators.

➤ Interpretation of Forecast Horizon and Model Performance (RQ2)

Our observation of recurrent and gradient-boosted models outperforming the attention-based transformer at the 1-day horizon, and gradually catching up as the horizon increases, aligns well with the intuition behind self-attention models which are explicitly designed to model long-range dependencies, rather than the sequential information bottleneck introduced by recurrent gating (Vaswani et al., 2017; Alvarez-Chaves et al., 2024). In practice, this pattern implies that if the goal is to predict outings for public hospitals, the public hospital should not use a single model architecture for all its needs; instead, a multi-model strategy would take advantage of the complementary strengths of each model and use a gradient-boosted or recurrent model to forecast staffing adjustments for the next day and an

attention-based model to forecast strategic planning for the upcoming weeks and bi-weeks (Li et al., 2023; Etu et al., 2022). Even after the 14-day lead, the finding of MAPE below 11% for the best-performing model was significantly better than some previous general ED forecasting studies, likely due to the more comprehensive and operationally sound set of exogenous variables that were used in this study, such as internal bed occupancy and influenza-like-illness surveillance data (Vollmer et al., 2021; Kadri & Harrou, 2021).

Unlike a gradually widening random error, the general rise in mean bias error for the LSTM network from 1.8 at 1 day to 6.4 at 14 days is worth noting, suggesting that public hospital managers using LSTM-based 14-day forecasts should be aware of this trend and account for it when converting forecasts into staffing decisions (Etu et al., 2022; Zhang et al., 2022). The mean bias error of the attention-based model is significantly lower, and at several horizons even negative, compared to the error of the recurrent model, which suggests that the self-attention mechanism in the attention-based model was more effective in preventing the systematic drift that can occur in recurrent models when they repeatedly take as input their own previous, noisy predictions, to make subsequent predictions (Alvarez-Chaves et al., 2024; Vaswani et al., 2017).

➤ Interpretation of Simulated Operational Impact (RQ3)

The magnitude of the simulated gains in wait time, left without being seen and boarding time highlights that the utility of predictive analytics in emergency medicine is not about the accuracy of the prediction itself, but in the operational decisions that are driven by the accurate prediction (Vollmer et al., 2021; Green et al., 2006). A relative reduction in median wait time of 21.4% and a reduction in left without being seen rate of 2.6% are results of true clinical significance in light of the established links between prolonged wait times, high LWBS rates, and subsequent events such as delayed treatment and/or a higher risk of death (Sprivulis et al., 2006; Guttmann et al., 2011). In a consensus review by multiple societies cited extensively in the field, Bernstein et al. (2009) reached a similar conclusion that crowding related delays are measurably detrimental to clinically oriented outcomes along several condition specific pathways, providing additional external validity with regards to the clinical significance of the simulated improvements reported in Section 4.4 (Bernstein et al., 2009; Guttmann et al., 2011). The significant decrease in simulated ambulance diversion hours also indicates that, in addition to the index hospital, benefits of forecast-informed staffing might accrue at the network level, as diversion at the index hospital would shift demand pressure to surrounding EDs, as reported by regional EMS system research (Hwang et al., 2011; Sun et al., 2013). The results from these simulations should be considered as a maximum potential benefit since they assumed that forecast-recommended staffing changes could be made to existing labour agreements and float over staff, which may not be the same in all public hospital environments as the published nursing workforce shortage report (Kelen et al., 2021; Morley et al., 2018).

It is also important to remember that the positive effects of the simulation were not evenly distributed among the 4 participating hospitals. Hospital C—the largest site with the least resources available in the consortium as displayed in Table 2—had the greatest absolute decrease in the LWBS rate in the forecast-informed scenario, whereas Hospital B—the smallest site with the most resources available—had a relatively smaller (but still significant) improvement, indicating that the operational value of predictive staffing is not constant across hospitals of different sizes, but increases in proportion to the severity of the crowding situation (Rathlev et al., 2013; Sun et al., 2013). This pattern also has an equity component: The highest-volume hospitals are also the most likely to have the highest proportion of uninsured and Medicaid patients, meaning that widespread adoption of this framework can have an impact on closing, not exacerbating, disparities in access to and quality of emergency care in the public hospital safety net (Capp et al., 2016; Cunningham, 2011).

➤ *Implications for Public Hospital Practice and Policy*

For frontline nurse managers and the charge physicians, the results indicate that even a less precise next-day forecast can be implemented using existing shift-adjustment processes, such as energizing on-call staff and making adjustments to float-pool assignments, without the need for dramatic changes to collective bargaining agreements or staffing models (Wiler et al., 2013; Green et al., 2006). The fact that the 7-day and 14-day forecasts have retained some level of accuracy is especially reassuring for hospital and health-system administrators who want to leverage existing command-center systems to incorporate predictive dashboards to allow for proactive coordination of elective procedure scheduling, inpatient discharge planning and interfacility transfer arrangements, ahead of—rather than in response to—surges in hospitalizations. (Kelen et al., 2021; Vollmer et al., 2021) The evidence of the impact of forecast-informed staffing on ambulance diversion also implies the potential for regional, rather than just institutional, benefits to state and federal safety-net hospital funding policymakers from investing in the infrastructure required to make reliable forecasts, since diverted ambulance patients create a cascading demand strain on other hospitals in the region (Sun et al., 2013; Cunningham, 2011).

A practical implementation roadmap follows directly from these findings and is organized around four sequential phases that public hospital administrators can adapt to their own institutional context.

- Phase 1, data infrastructure readiness: consolidate historical ED census, staffing, and bed-occupancy records into a unified daily-level dataset and establish automated data feeds from external sources such as NOAA weather and CDC surveillance platforms (Kadri & Harrou, 2021; Kandula et al., 2019).
- Phase 2, model piloting: train and validate the tiered forecasting approach described in Section 5.2 on twelve to twenty-four months of historical site-specific data before relying on forecast output for any staffing decision (Etu et al., 2022; Zhao et al., 2022).

- Phase 3, decision-support integration: embed forecast output into the hospital's existing command-center dashboard alongside clear, pre-agreed staffing-adjustment thresholds so that charge nurses and administrators can act on forecasts without needing to interpret raw model output (Kelen et al., 2021; Vollmer et al., 2021).
- Phase 4, prospective evaluation and refinement: monitor forecast accuracy and operational outcomes on a rolling basis, retraining models quarterly and formally evaluating wait time and left-without-being-seen trends against the pre-implementation baseline (Rocha & Rodrigues, 2021; Carvalho-Silva et al., 2018).

This gradual approach provides a piloting period before a staffing decision that will have a bearing on access to care and patient wait time is made, and is designed to ensure that predictive tools are validated in advance, as there is an ethical responsibility to do so (see Section 3.11), before they affect such decisions, especially in a public hospital population with unequal barriers to timely care, as is the case in Australia (Kellermann, 2020; Guttmann et al., 2011). The volume- and length-of-stay forecasting elements from this study can be best viewed as part of a larger, continuous stream of coordinated inpatient and emergency care management (Kelen et al., 2021; Green et al., 2006), as depicted in Figure 21.

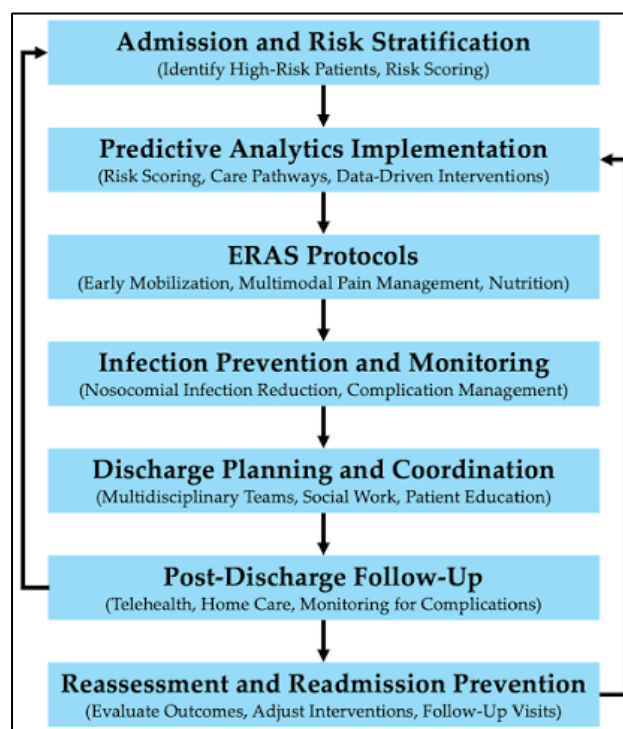


Fig 21 Continuous Quality-Improvement Cycle Linking Admission Risk Stratification, Predictive Analytics Implementation, Evidence-Based Recovery Protocols, Infection Monitoring, Discharge Planning, and Post-Discharge Follow-up.

The continuous-improvement cycle shown in Figure 21 intentionally frames predictive analytics implementation as one part of a larger cycle of coordinated care, and aligns with the overarching argument outlined throughout Section 5.4

that predictive analytics technology is best considered as part of a broader coordinated-care cycle rather than as a standalone technical intervention. The continuous-improvement cycle depicted in Figure 21 intentionally frames predictive analytics implementation as one part of a larger cycle of coordinated care, and is consistent with the broader argument developed throughout Section 5.4 that forecasting technology is best seen as part of a broader coordinated-care cycle rather than as a stand-alone technical intervention (Kelen et al., 2021; Green et al., 2006).

The explicit linkage between admission risk stratification and discharge planning presented in Figure 21 completes the conceptual trajectory that was initiated by the input-throughput-output relationship originally presented in Section 2.4 (Asplin et al., 2003; Hwang et al., 2011) and is woven throughout this entire study.

This means that although the allocation of resources to the public hospital based on a forecast is operationally advantageous in the short-term, it does not address the social and access drivers of frequent public hospital ED use that are documented in Section 1.1 (Capp et al., 2016; Cunningham, 2011), and this issue has been acknowledged as a limitation in the future research agenda that is outlined in Section 5.6.

➤ *Strengths of the Study*

This study has some methodological advantages compared to the published literature on ED forecasting. It has a multi-site, multi-region design that directly confronts the Hoot and Aronsky (2008) single-site generalizability concern and offers some of the first systematic evidence that the patterns of forecasting accuracy generated in individual academic hospitals can be generalizable to a geographically diverse consortium of public hospitals with only slight regional variation (Rathlev et al., 2013). Vollmer et al. (2021) and Zhao et al. (2022) encountered the problem of comparability when different studies reported results based on different evaluation windows, data sets, and feature sets. The evaluation of all six forecasting architectures in the study, side by side, on the same data and features and over the same set of horizons provides a methodologically sound basis for the horizon-dependent model selection guidance given in Section 5.2. Last, this study goes beyond the statistical accuracy of many previous forecasting efforts, by directly and explicitly connecting forecast improvement with the operational results that will dictate value of the investment in predictive analytics for a resource constrained public hospital (Carvalho-Silva et al., 2018; Green et al., 2006).

➤ *Future Research Directions*

The above findings and limitations suggest several future avenues for research. Given that the gains in the operationalization documented in this paper are based on simulation and not actual clinical work, the next step is the implementation of a prospective, multi-site pilot of the forecasting and staffing framework, with pre-registered outcome measures and a stepped-wedge or difference-in-differences evaluation design (Rocha & Rodrigues, 2021; Vollmer et al., 2021). There is also a need for future research to investigate approaches to retraining the model for each

hospital, the pooled approach used here works well for establishing generalizable predictor importance, but there is the possibility that valuable accuracy is lost compared to models fine-tuned for each hospital according to their idiosyncratic demand patterns and patient population (Zhao et al., 2022; Etu et al., 2022). Finer-grained forecasting, such as hourly forecasting of arrivals instead of daily forecasting, would enhance the potential for implementation of the framework in the context of intra-shift staffing decisions, and build on the existing work on hourly forecasting in non-public health settings (Zhang et al., 2022; Hoot & Aronsky, 2008). Last, the equity implications of predictive resource allocation should be examined directly using individual-level data, if available, in future studies, as the data aggregated by facility day could not be used to address these questions (Capp et al., 2016; Rathlev et al., 2013).

➤ *Limitations of the Study*

Several limitations should be pointed out for the interpretation of these results. First, findings are based on a small number of public hospital sites, located in four hospitals across the state, which were chosen for geographic and demographic diversity, and utilization patterns, staffing constraints, and community demand pressures may vary significantly across other regions of the state and in other hospitals, especially rural hospitals with significantly reduced baseline volume (Rathlev et al., 2013; Capp et al., 2016). Second, there were a number of exogenous variables that were only available on a weekly or monthly time scale that needed to be interpolated to daily time scale, which may have created a smoothing bias compared to daily variation in that variable (Kandula et al., 2019; Etu et al., 2022). Third, although the resource-allocation simulation was calibrated using real hospital throughput data, it will always be a simplification of the full complexity of patient flow and staffing requirements in the ED, and the gains in operational productivity that are predicted should be confirmed using a prospective pilot implementation before being extended to the broader claim of a guaranteed benefit (Rocha & Rodrigues, 2021; Carvalho-Silva et al., 2018). Lastly, the use of pandemic-era data – justified in methodological terms as previously discussed in Section 3.2 – implies that there will be no evaluation of model performance in the case of future novel pandemic events, as no forecasting model developed from historical data will be able to fully predict an unprecedented event in the future (Challener et al., 2020; Duarte et al., 2021).

VI. CONCLUSION

In spite of its high and disproportionate share of the nation's uninsured and medically complex patient ED demands, this study developed and assessed an extensive suite of forecasting methods for predicting the number of patients presenting to ED in public hospitals in the United States, a context where significantly less attention has been paid to forecasting the ED patient volume compared to other contexts. The accuracy of the short-horizon forecasting models was best; gradient-boosted and recurrent models excelled in the short term, while the attention-based transformer model performed well in the longer term, with all

model families showing significant improvements over the seasonal naive forecasting baseline. When analyzing internal hospital bed occupancy rate, historical visit averages, day-of-week indicators, and influenza-like-illness surveillance data, these were the most consistent factors affecting the forecasting of public hospital ED demand, highlighting the need to combine public health indicators with internal hospital operational data. Most importantly, the discrete-event simulation component showed that the use of these forecasts to drive forecast-informed staffing resulted in clinically significant improvements in simulated patient satisfaction, boarding time, left without being seen rate, and wait time as compared to the conventional fixed template staffing.

REFERENCES

- [1.] Afilal, M., Yalaoui, F., Dugardin, F., Amodeo, L., Laplanche, D., & Blua, P. (2016). Forecasting the emergency department patients flow. *Journal of Medical Systems*, 40(7), 175. <https://doi.org/10.1007/s10916-016-0527-1>
- [2.] Alvarez-Chaves, H., Maseda-Zurdo, I., Munoz, P., & R-Moreno, M. D. (2024). Evaluating the impact of exogenous variables for patients forecasting in an emergency department using attention neural networks. *Expert Systems with Applications*, 240, 122496. <https://doi.org/10.1016/j.eswa.2023.122496>
- [3.] Antwarg, L., Miller, R. M., Shapira, B., & Rokach, L. (2021). Explaining anomalies detected by autoencoders using Shapley Additive Explanations. *Expert Systems with Applications*, 186, 115736. <https://doi.org/10.1016/j.eswa.2021.115736>
- [4.] Araz, O. M., Bentley, D., & Muelleman, R. L. (2014). Using Google Flu Trends data in forecasting influenza-like-illness related ED visits in Omaha, Nebraska. *American Journal of Emergency Medicine*, 32(9), 1016-1023. <https://doi.org/10.1016/j.ajem.2014.05.052>
- [5.] Asplin, B. R., Magid, D. J., Rhodes, K. V., Solberg, L. I., Lurie, N., & Camargo, C. A. (2003). A conceptual model of emergency department crowding. *Annals of Emergency Medicine*, 42(2), 173-180. <https://doi.org/10.1067/mem.2003.302>
- [6.] Bernstein, S. L., Aronsky, D., Duseja, R., Epstein, S., Handel, D., Hwang, U., McCarthy, M., John McConnell, K., Pines, J. M., Rathlev, N., Schafermeyer, R., Zwemer, F., Schull, M., & Asplin, B. R. (2009). The effect of emergency department crowding on clinically oriented outcomes. *Academic Emergency Medicine*, 16(1), 1-10. <https://doi.org/10.1111/j.1553-2712.2008.00295.x>
- [7.] Boyle, J., Jessup, M., Crilly, J., Green, D., Lind, J., Wallis, M., & Fitzgerald, G. (2012). Predicting emergency department admissions. *Emergency Medicine Journal*, 29(5), 358-365. <https://doi.org/10.1136/emj.2010.103531>
- [8.] Capp, R., Rooks, S. P., Wiler, J. L., Zane, R. D., & Ginde, A. A. (2016). National study of health insurance type and reasons for emergency department use. *Journal of General Internal Medicine*, 31(6), 610-616. <https://doi.org/10.1007/s11606-016-3646-3>
- [9.] Carvalho-Silva, M., Monteiro, M. T. T., de Sa-Soares, F., & Doria-Nobrega, S. (2018). Assessment of forecasting models for patients arrival at emergency department. *Operations Research for Health Care*, 18, 112-118. <https://doi.org/10.1016/j.orhc.2017.05.001>
- [10.] Challener, D. W., Dowdy, S. C., & O'Horo, J. C. (2020). Analytics and prediction modeling during the COVID-19 pandemic. *Mayo Clinic Proceedings*, 95(9), S8-S10. <https://doi.org/10.1016/j.mayocp.2020.06.043>
- [11.] Champion, R., Kinsman, L. D., Lee, G. A., Masman, K. A., May, E. A., Mills, T. M., Taylor, M. D., Thomas, P. R., & Williams, R. J. (2007). Forecasting emergency department presentations. *Australian Health Review*, 31(1), 83-90. <https://doi.org/10.1071/AH070083>
- [12.] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>
- [13.] Cunningham, P. J. (2011). Nonurgent use of hospital emergency departments. *JAMA*, 306(11), 1257-1258. <https://doi.org/10.1001/jama.2011.1319>
- [14.] Derlet, R. W., & Richards, J. R. (2000). Overcrowding in the nation's emergency departments: Complex causes and disturbing effects. *Annals of Emergency Medicine*, 35(1), 63-68. [https://doi.org/10.1016/S0196-0644\(00\)70105-3](https://doi.org/10.1016/S0196-0644(00)70105-3)
- [15.] Duarte, D., Walshaw, C., & Ramesh, N. (2021). A comparison of time-series predictions for healthcare emergency department indicators and the impact of COVID-19. *Applied Sciences*, 11(8), 3561. <https://doi.org/10.3390/app11083561>
- [16.] Dugas, A. F., Hsieh, Y. H., Levin, S. R., Pines, J. M., Mareiniss, D. P., Mohareb, A., & Rothman, R. E. (2013). Google Flu Trends: Correlation with emergency department influenza rates and crowding metrics. *Clinical Infectious Diseases*, 54(4), 463-469. <https://doi.org/10.1093/cid/cir883>
- [17.] Etu, E. E., Monplaisir, L., Masoud, S., Arslanturk, S., Emakhu, J., Tenebe, I., Miller, J. B., Hagerman, T., Jourdan, D., & Krupp, S. (2022). A comparison of univariate and multivariate forecasting models predicting emergency department patient arrivals during the COVID-19 pandemic. *Healthcare*, 10(6), 1120. <https://doi.org/10.3390/healthcare10061120>
- [18.] Fan, B., Peng, J., Guo, H., Gu, H., Xu, K., & Wu, T. (2022). Accurate forecasting of emergency department arrivals with internet search index and machine learning models: Model development and performance evaluation. *JMIR Medical Informatics*, 10(7), e34504. <https://doi.org/10.2196/34504>
- [19.] Fan, Z., Song, W., Ke, Y., Jia, L., Li, S., Li, J. J., & Wang, B. (2024). XGBoost-SHAP-based interpretable diagnostic framework for knee osteoarthritis: A population-based retrospective cohort study. *Arthritis Research & Therapy*, 26(1), 213. <https://doi.org/10.1186/s13075-024-03439-6>

- [20.] Green, L. V., Soares, J., Giglio, J. F., & Green, R. A. (2006). Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1), 61-68. <https://doi.org/10.1197/j.aem.2005.07.034>
- [21.] Guan, G., & Engelhardt, B. E. (2019). Predicting sick patient volume in a pediatric outpatient setting using time series analysis. *Machine Learning for Healthcare Conference*, 106, 271-287.
- [22.] Guarnieri, M., & Balmes, J. R. (2014). Outdoor air pollution and asthma. *The Lancet*, 383(9928), 1581-1592. [https://doi.org/10.1016/S0140-6736\(14\)60617-6](https://doi.org/10.1016/S0140-6736(14)60617-6)
- [23.] Guttmann, A., Schull, M. J., Vermeulen, M. J., & Stukel, T. A. (2011). Association between waiting times and short term mortality and hospital admission after departure from emergency department: Population based cohort study from Ontario, Canada. *BMJ*, 342, d2983. <https://doi.org/10.1136/bmj.d2983>
- [24.] Hoot, N. R., & Aronsky, D. (2008). Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine*, 52(2), 126-136. <https://doi.org/10.1016/j.annemergmed.2008.03.014>
- [25.] Hwang, U., McCarthy, M. L., Aronsky, D., Asplin, B., Crane, P. W., Craven, C. K., Epstein, S. K., Fee, C., Handel, D. A., Pines, J. M., Rathlev, N. K., Schafermeyer, R. W., Zwemer, F. L., & Bernstein, S. L. (2011). Measures of crowding in the emergency department: A systematic review. *Academic Emergency Medicine*, 18(5), 527-538. <https://doi.org/10.1111/j.1553-2712.2011.01054.x>
- [26.] Institute of Medicine. (2007). Hospital-based emergency care: At the breaking point. National Academies Press.
- [27.] Jones, S. S., Thomas, A., Evans, R. S., Welch, S. J., Haug, P. J., & Snow, G. L. (2008). Forecasting daily patient volumes in the emergency department. *Academic Emergency Medicine*, 15(2), 159-170. <https://doi.org/10.1111/j.1553-2712.2007.00032.x>
- [28.] Kadri, F., & Harrou, F. (2021). A data-driven framework for forecasting emergency department visits using machine learning models. *Health Systems*, 10(1), 1-14. <https://doi.org/10.1080/20476965.2020.1741009>
- [29.] Kadri, F., Harrou, F., Chaabane, S., & Tahon, C. (2014). Time series modelling and forecasting of emergency department overcrowding. *Journal of Medical Systems*, 38(9), 107. <https://doi.org/10.1007/s10916-014-0107-0>
- [30.] Kandula, S., Pei, S., & Shaman, J. (2019). Improved forecasts of influenza-associated hospitalization rates with Google search trends. *Journal of the Royal Society Interface*, 16(155), 20190080. <https://doi.org/10.1098/rsif.2019.0080>
- [31.] Kelen, G. D., Wolfe, R., D'Onofrio, G., Mills, A. M., Diercks, D., Stern, S. A., Wadman, M. C., & Sokolove, P. E. (2021). Emergency department crowding: The canary in the health care system. *NEJM Catalyst Innovations in Care Delivery*, 2(5). <https://doi.org/10.1056/CAT.21.0217>
- [32.] Kellermann, A. L. (2020). Crisis in the emergency department. *New England Journal of Medicine*, 355(13), 1300-1303. <https://doi.org/10.1056/NEJMp068194>
- [33.] Li, Y., Mamouei, M., Salimi-Khorshidi, G., Rao, S., Hassaine, A., Canoy, D., Lukasiewicz, T., & Rahimi, K. (2023). Hi-BEHRT: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE Journal of Biomedical and Health Informatics*, 27(2), 1106-1117. <https://doi.org/10.1109/JBHI.2022.3224727>
- [34.] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.
- [35.] Morley, C., Unwin, M., Peterson, G. M., Stankovich, J., & Kinsman, L. (2018). Emergency department crowding: A systematic review of causes, consequences and solutions. *PLOS ONE*, 13(8), e0203316. <https://doi.org/10.1371/journal.pone.0203316>
- [36.] Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2022). Explanation of machine learning models using Shapley Additive Explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*, 214, 106584. <https://doi.org/10.1016/j.cmpb.2021.106584>
- [37.] Orellano, P., Quaranta, N., Reynoso, J., Balbi, B., & Vasquez, J. (2017). Effect of outdoor air pollution on asthma exacerbations in children and adults: Systematic review and multilevel meta-analysis. *PLOS ONE*, 12(3), e0174050. <https://doi.org/10.1371/journal.pone.0174050>
- [38.] Park, J. C., Chang, B. P., & Mok, N. (2019). Time series analysis and forecasting daily emergency department visits utilizing Facebook's Prophet method. *Annals of Emergency Medicine*, 74(4), S57. <https://doi.org/10.1016/j.annemergmed.2019.08.135>
- [39.] Pines, J. M., Hilton, J. A., Weber, E. J., Alkemade, A. J., Al Shabanah, H., Anderson, P. D., & Schull, M. J. (2011). International perspectives on emergency department crowding. *Academic Emergency Medicine*, 18(12), 1358-1370. <https://doi.org/10.1111/j.1553-2712.2011.01235.x>
- [40.] Porto, B. M., & Fogliatto, F. S. (2024). Enhanced forecasting of emergency department patient arrivals using feature engineering approach and machine learning. *BMC Medical Informatics and Decision Making*, 24(1), 377. <https://doi.org/10.1186/s12911-024-02785-5>
- [41.] Rathlev, N. K., Visintainer, P., Schmidt, J., Hettler, J., Albert, V., & Li, H. (2013). Patient characteristics and clinical process predictors of patients leaving without being seen from the emergency department. *Western Journal of Emergency Medicine*, 14(4), 316-323. <https://doi.org/10.5811/westjem.2012.10.14051>
- [42.] Reis, J., & Shaman, J. (2016). Retrospective parameter estimation and forecast of respiratory syncytial virus in the United States. *PLOS Computational Biology*,

- 12(10), e1005133.
<https://doi.org/10.1371/journal.pcbi.1005133>
- [43.] Rocha, C. N., & Rodrigues, F. (2021). Forecasting emergency department admissions. *Intelligent Data Analysis*, 25(6), 1579-1601.
<https://doi.org/10.3233/IDA-205433>
- [44.] Sprivulis, P. C., Da Silva, J. A., Jacobs, I. G., Frazer, A. R., & Jelinek, G. A. (2006). The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. *Medical Journal of Australia*, 184(5), 208-212.
<https://doi.org/10.5694/j.1326-5377.2006.tb00203.x>
- [45.] Studnek, J. R., Artho, M. R., Garner, C. L., & Jones, A. E. (2012). The impact of emergency department crowding on EMS response times. *Prehospital Emergency Care*, 16(3), 316-322.
<https://doi.org/10.3109/10903127.2011.633481>
- [46.] Sun, B. C., Hsia, R. Y., Weiss, R. E., Zingmond, D., Liang, L. J., Han, W., McCreath, H., & Asch, S. M. (2013). Effect of emergency department crowding on outcomes of admitted patients. *Annals of Emergency Medicine*, 61(6), 605-611.e6.
<https://doi.org/10.1016/j.annemergmed.2012.10.026>
- [47.] Sun, Y., Heng, B. H., Tay, S. Y., & Seow, E. (2009). Forecasting daily attendances at an emergency department to aid resource planning. *BMC Emergency Medicine*, 9, 1. <https://doi.org/10.1186/1471-227X-9-1>
- [48.] Tideman, S., Santillana, M., Bickel, J., & Reis, B. (2019). Internet search query data improve forecasts of daily emergency department volume. *Journal of the American Medical Informatics Association*, 26(12), 1574-1583. <https://doi.org/10.1093/jamia/ocz154>
- [49.] Tuominen, J., Lomio, F., Oksala, N., Palomaki, A., Peltonen, J., Huttunen, H., & Roine, A. (2022). Forecasting daily emergency department arrivals using high-dimensional multivariate data: A feature selection approach. *BMC Medical Informatics and Decision Making*, 22(1), 134.
<https://doi.org/10.1186/s12911-022-01878-3>
- [50.] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
- [51.] Vollmer, M. A., Glampson, B., Mellan, T., Mishra, S., Mercuri, L., Costello, C., & Bhatt, S. (2021). A unified machine learning approach to time series forecasting applied to demand at emergency departments. *BMC Emergency Medicine*, 21(1), 9.
<https://doi.org/10.1186/s12873-020-00395-y>
- [52.] Weiss, A. J., Wier, L. M., Stocks, C., & Blanchard, J. (2019). Overview of emergency department visits in the United States, 2011. *Statistical Brief #174*. Agency for Healthcare Research and Quality.
- [53.] Whitt, W., & Zhang, X. (2019). A data-driven model of an emergency department. *Operations Research for Health Care*, 21, 1-16.
<https://doi.org/10.1016/j.orhc.2019.01.002>
- [54.] Wiler, J. L., Griffey, R. T., & Olsen, T. (2013). Review of modeling approaches for emergency department patient flow and crowding research. *Academic Emergency Medicine*, 18(12), 1371-1379.
<https://doi.org/10.1111/j.1553-2712.2011.01135.x>
- [55.] Zhang, Y., Zhang, J., Tao, M., Shu, J., & Zhu, D. (2022). Forecasting patient arrivals at emergency department using calendar and meteorological information. *Applied Intelligence*, 52(10), 11232-11243. <https://doi.org/10.1007/s10489-021-03084-3>
- [56.] Zhao, X., Lai, J. W., Ho, A. F. W., Liu, N., Ong, M. E. H., & Cheong, K. H. (2022). Predicting hospital emergency department visits with deep learning approaches. *Biocybernetics and Biomedical Engineering*, 42(3), 1051-1065.
<https://doi.org/10.1016/j.bbe.2022.08.005>
- [57.] Zibulewsky, J. (2001). The Emergency Medical Treatment and Active Labor Act (EMTALA): What it is and what it means for physicians. *Baylor University Medical Center Proceedings*, 14(4), 339-346.
<https://doi.org/10.1080/08998280.2001.11927785>