Occupancy Prediction with Patient Data: Evaluating Time-Series, Patient-Level Aggregation, and Deep Set Models

Song-Hee Kim
SNU Business School, Seoul National University, Seoul, South Korea, songheekim@snu.ac.kr

William Overman
Graduate School of Business, Stanford University, Stanford, CA, United States, wpo@stanford.edu

Jean Pauphilet
Management Science and Operations, London Business School, London, United Kingdom, jpauphilet@london.edu

Won Chul Cha
Department of Emergency Medicine, Samsung Medical Center, Seoul, South Korea, wc.cha@samsung.com

Problem definition: With the growing availability of patient-level data and advanced data analytics tools, hospitals are increasingly developing prediction models to improve operational efficiency. This study evaluates three modeling approaches for predicting Emergency Department (ED) occupancy using patient-level information: time-series models enhanced with summary statistics of current ED patients, bottom-up models that aggregate individual patient-level predictions, and deep set neural networks that predict occupancy directly from the concatenation of all patient-level information available. **Methodology/Results:** Our results show that time-series models, despite their simplicity and relatively low complexity, significantly benefit from the inclusion of summary information about current ED patients, leading to a 15-30% reduction in Mean Absolute Error (MAE). Bottom-up approaches, offering dual-level (patient and ED) predictions, not only improve interpretability, but also significantly improve occupancy prediction accuracy by 35% compared to time-series models. However, to achieve these gains, we show that bottom-up pipelines require an additional calibration step (and careful re-calibration over time). Meanwhile, deep set models show promise with performance between that of time-series and bottom-up models. However, their performance on low-volume days suggests limitations for smaller institutions. We assess the robustness of our findings across different prediction horizons and validate our results using data from 10 additional hospitals. **Managerial implications:** This study demonstrates the value of leveraging patient-level data in improving ED occupancy prediction and highlights the importance of model selection, calibration, and fine-tuning when using patient-level data to predict ED occupancy.

Key words: Patient-level prediction, machine learning, deep learning, calibration, bias, nonstationarity, distribution shift, ED occupancy, healthcare

1. Introduction

To improve their operations, hospitals have been developing and deploying analytics tools to predict key indicators of future medical needs and supplies, such as future arrivals to the Emergency Department (ED; Wargon et al. 2009), daily inpatient admissions (Zhou et al. 2018), discharges (McCoy et al. 2018), and medical bookings (Piccialli et al. 2021). Time-series models are particularly relevant and accurate for estimating such system-level quantities. In recent year, these models have benefited from advancements in nonlinear machine learning methods, including Meta's Prophet toolbox (see McCoy et al. 2018, for an example on discharge volume prediction), general regression neural networks (see Duarte et al. 2021, for an application to ED volume forecasting), and long short-term memory models (see Borges and Nascimento 2022, in the context of COVID-19). Many studies have also investigated the value of external data, such as weather variables (see Wargon et al. 2009, for a review) or internet search trends (Tideman et al. 2019, Trevino et al. 2022, Fan et al. 2022), in making more accurate predictions for system-level quantities.

Concurrently, the increasing availability of granular and multimodal data in electronic health records (EHRs) creates opportunities to predict a myriad of patient-level outcomes, including length of stay, mortality, and readmissions (see, e.g., Awad et al. 2017, Xiao et al. 2018, Stone et al. 2022, Bacchi et al. 2022, for reviews). Potentially, these patient-level predictions can be aggregated to make even more accurate system-wide forecasts, which, in turn, can be inform better decisions regarding capacity dimensionning, staff and appointment scheduling, and dynamic resource allocation. For example, recent studies have aggregated patient-level length of stay predictions to predict unit-level occupancy (Bertsimas et al. 2021, Wang et al. 2022, King et al. 2022).

However, analyzing raw EHR data is complicated due to data quality issues (Feder 2018), the presence of missing information (Kharrazi et al. 2014), and the prevalence of unstructured data (Tayefi et al. 2021). In addition, integrating analytics-based approaches into real-life decision-making introduces further challenges related to real-time data extraction and processing, communication and interpretation of analytical results, and adoption of the tool by practitioners, which become more significant as the complexity of the data and models increases.

Furthermore, recent studies that aggregate patient-level length of stay predictions to forecast future hospital occupancy (e.g., Bertsimas et al. 2021, Wang et al. 2022, King et al. 2022) report estimates that appear to be biased[1]. This suggests that aggregating patient-level predictions for overall occupancy estimation may introduce non-trivial issues that have been overlooked.

Alternatively, recent models in the deep learning literature, such as the deep sets of Zaheer et al. (2017), are designed to handle size-varying inputs, such as concatenating patient-level information

[1] For example: "The prediction pipeline underestimated the number of admissions within the 8-hour window but performed better than the benchmark." (King et al. 2022)

for patients currently present in the hospital or unit. These models could predict system-level quantities like occupancy or discharges directly, leveraging the rich information contained in patient-level data. To the best of our knowledge, however, this methodology is not widely known, let alone used, in the healthcare operations literature.

Given that many important operational decisions are made at a less granular level, for which time-series models are well suited, one might wonder *whether the benefits of using patient-level data outweigh the increased complexity*. In this paper, we use the prediction of occupancy in the emergency department (ED) as an example to investigate whether and how patient-level information can enhance predictions on system-wide quantities. Our goal is to shed light on the opportunities and challenges of working with patient-level information and provide a practical roadmap for analytics teams in healthcare organizations considering sophisticated patient-level predictions to improve their system operations.

We have access to data from the ED of a high-volume academic hospital and aim to predict short-term occupancy. We evaluate three models: a *time-series* model; a *bottom-up* model that predicts the remaining length of stay for each patient in the ED and aggregates these predictions; and a *direct* model using deep set neural networks (Zaheer et al. 2017) to predict future occupancy directly from the concatenation of all patient information. We present our data and introduce our three models in Section 2. Additionally, we analytically explain the emergence of systematic bias in the bottom-up model, as experienced in the literature (Bertsimas et al. 2021, Wang et al. 2022, King et al. 2022), and propose effective solutions to correct it. At a high-level, this bias comes from a misalignment between the patient- and system-level prediction tasks, leading to a sort of inspection paradox (Stein and Dattero 1985).

We then study each model separately. In Section 3, we demonstrate the power of time-series models, especially nonlinear models and those incorporating summary information about current ED patients. In Section 4, we implement bottom-up models, confirming the bias resulting from the misalignment between the patient-level predictive task and the downstream system-level task, and illustrating the effectiveness of the calibration steps we proposed. In Section 5, we develop a deep set architecture to predict ED occupancy directly and investigate the nature and relevance of the patient embedding learned with this approach.

Finally, in Section 6, we compare the performance of the time-series, bottom-up, and direct models on our test period. As many healthcare (and real-world) environments, our setting suffers from strong nonstationarities. We discuss the challenges in accurately detecting distributional shifts in practice, especially when using patient-level data, and propose simple retraining strategies to (partially) alleviate the decalibration of models over time. We assess the robustness of our findings in Section 7 by evaluating the effect of the prediction horizon and replicating our analysis in 10 additional hospitals.

Table 1 Prediction horizons typically used in short-term ED occupancy prediction.

Study	Prediction horizon
Zeltyn et al. (2011)	8 hours
Schweigler et al. (2009)	4 and 12 hours
Jones et al. (2009)	1 to 24 hours
Whitt and Zhang (2019)	1 to 6 hours
Cheng et al. (2021)	1 to 4 hours
King et al. (2022)	4 hours
Tuominen et al. (2023)	1 to 24 hours
Tuominen et al. (2024)	24 hours

2. Problem Description and Modeling

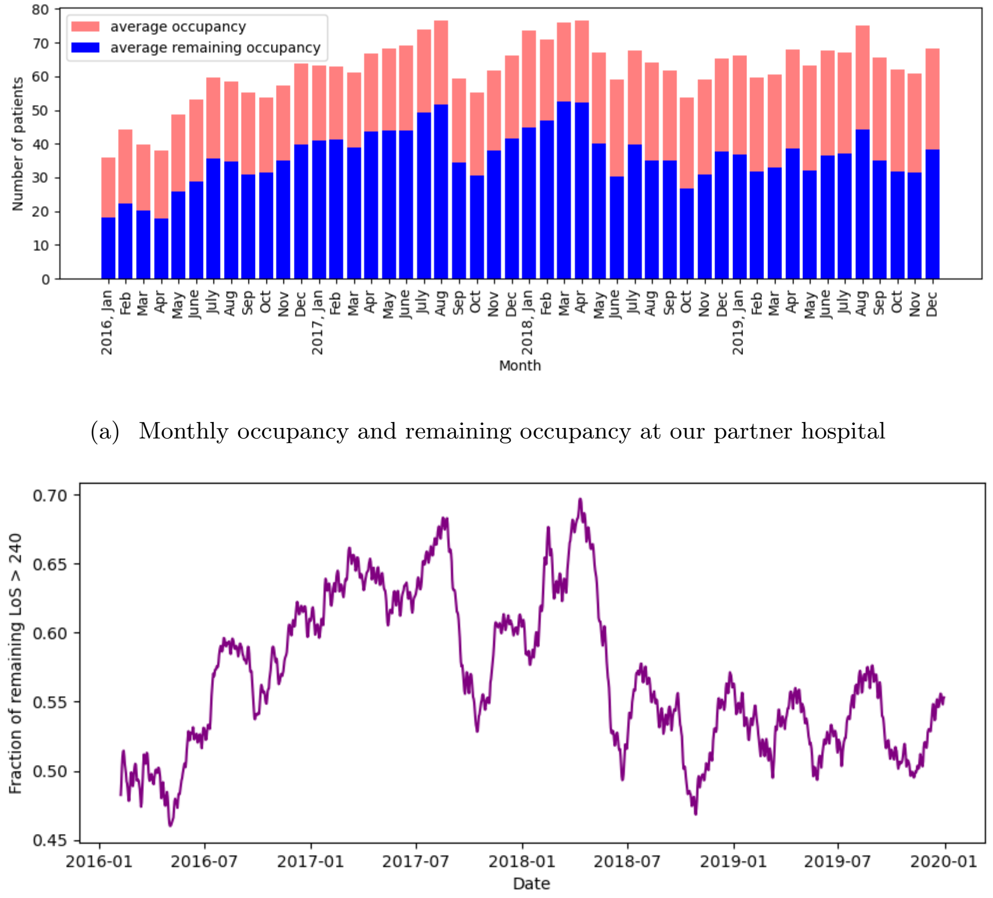
In this section, we first describe our ED occupancy problem in Section 2.1 and the data we have access to in Section 2.2. Then, in Section 2.3, we introduce the three families of models that can be (and, for some of them, have been) used to address this problem. In Section 2.4, we conclude this section by analyzing the emergence of bias in bottom-up approaches and propose simple calibration steps to correct it.

2.1. Case Study: Short-Term ED Occupancy Prediction

Predicting short-term occupancy in the ED is an important question for daily operations, especially to identify needs for surge staffing (see, e.g., Hu et al. 2021). Typically, 'short-term' for ED operations mean 'within the next hours'. Indeed, several studies in the literature have been interested in predicting future ED occupancy with varying prediction horizons (ranging from one hour to one day; see Table 1), in order to inform operational decisions like ambulance diversion (Lagoe and Jastremski 1990, Lin et al. 2015, Katayama et al. 2017), surge staffing (Zeltyn et al. 2011, Hu et al. 2023), bed planning (King et al. 2022), or early warning systems (Tuominen et al. 2023). In our main study, we restrict our attention to predicting occupancy four hours ahead, but confirm the robustness of our findings to the prediction horizon in Section 7.1.

At a given point in time, total future occupancy can be decomposed into two components. First, the number of patients who are currently in the ED and that will still be in the ED in four hours. Second, patients who might arrive to the ED within the next four hours and stay sufficiently long to still be here in four hours. In this paper, since we are interested in developing and aggregating patient-level predictions, we focus our attention on the first component, namely the contribution of current ED patients to the 4-hour ED occupancy, which we will refer to as 'remaining occupancy' in the rest of the paper. Note that the second component can be estimated separately, e.g., using Poisson regression (see King et al. 2022, for an example in a healthcare setting).

To illustrate the complexity of predicting short-term occupancy, Figure 1(a) shows the daily occupancy and 'remaining occupancy' at our partner hospital throughout the study period, with



(a) Monthly occupancy and remaining occupancy at our partner hospital

(b) Rolling 30-day average of the fraction of current ED patients staying at least four hours

Figure 1 Average occupancy and average remaining occupancy in a high-volume academic hospital in Korea in 2016–2019.

averages calculated for each month. From this, we observe that the remaining occupancy fluctuates over time, mirroring fluctuations in occupancy. However, Figure 1(b), which displays the ratio of remaining occupancy to current occupancy, shows that the remaining occupancy cannot be simply predicted as a constant fraction of the current occupancy. The significant variation in the ratio of remaining to current occupancy highlights the challenge of our prediction task.

2.2. Data

We use data from a high-volume academic hospital in South Korea. Our data consists of all 306,317 ED visits between January 9, 2016 and December 31, 2019. We exclude 8 patients for whom the recorded arrival or departure time was missing, so our final dataset comprises 306,309 patients. We apply no further exclusion criteria, such as missing data or extreme length of stay.

Most of the patient-level information is available at triage, shortly after the patient arrives at the ED. Our data include demographic information about the patient (age and sex). In addition, several clinical features are obtained at triage: measurements of systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse rate (PR), respiration rate (RR), body temperature (Temp), and oxygen saturation (SpO2). Finally, South Korea has implemented a scoring system called the Korean Triage and Acuity Score (KTAS) which assigns a score from 1 (resuscitation) to 5 (non-urgent) depending on the patient's urgency (Kwon et al. 2018). Despite variability in the label and definition of each category, scoring systems are a common practice for patient triage in many countries. Beyond South Korea, the Emergency Severity Index in the United States, the Australasian Triage Scale, or the Canadian Triage and Acuity Scale, for example, implement a similar five-level scoring system for ED patients (see Gilboy et al. 2005, chapter 2). In our data, we observe that all these clinical features (vital sign measurements and KTAS score) are recorded within 6 minutes of the patient's arrival to the ED on average, with a standard deviation of 6.8 minutes. Given that this information is available shortly after arrival, we consider it as available upon arrival and use it to base our patient-level predictions. For simplicity, we restrict our attention to the data collected upon the patient's arrival and do not update it as the patient stays.[2] We do so also to be consistent with our robustness analysis in Section 7.2, where we us a national database to replicate our analysis to multiple hospitals and do not have access to dynamic variables. See Online Supplement EC.3 for summary statistics for ED length of stay (LoS) and all patient-level features.

Throughout our study, for all our predictive tasks, we treat the data from January 9, 2016 to December 31, 2018 as the training set, and the data from January 1, 2019 to December 31, 2019 as the test set. See Online Supplement EC.1 for a detailed description of the machine learning models and performance metrics used.

2.3. Modeling

Consider N patients observed over T time periods. For each patient $i = 1, \dots, N$ and each time period $t = 1, \dots, T$, $\delta_{i,t} \in \{0, 1\}$ indicates whether patient i is present in the ED at time t.

At time t, we denote $N_t := \sum_{i=1}^{N} \delta_{i,t}$ the number of patients in the ED. We are interested in predicting the 'remaining occupancy', i.e., how many of these N_t patients will still be in the ED at time $t+1$. Formally, we are interested in

$$z_{t+1} = \sum_{i:\delta_{i,t}=1} \delta_{i,t+1}. \tag{1}$$

Again, we emphasize that z_{t+1} does not correspond to the ED occupancy at time $t+1$, but only to the number of patients that were in the ED at time t and stayed until $t+1$. In particular, we

[2] Integrating time-varying values of vital signs or other laboratory or examination results should presumably improve the performance of patient-level models, at the expense of a more sophisticated software integration (e.g., triage data are often stored in a different database than the data collected during the rest of the ED visit).

have $z_{t+1} \leq N_t$ and $z_{t+1} \leq N_{t+1}$. With these notations, Figure 1(a) represents the monthly average of N_t (pink bars) and z_{t+1} (blue bars) in our dataset, and Figure 1(b) z_{t+1}/N_t. To estimate z_{t+1}, we need to estimate individual survival probabilities, $\mathbb{P}(\delta_{i,t+1} = 1 | \delta_{i,t} = 1)$.

2.3.1. Time-Series Models Time-series approaches implicitly assume that survival probabilities do not depend on patient characteristics but only on time-related features, leading to estimates of the form

$$\hat{z}_{t+1}^{\texttt{ts}} = N_t \times \varphi(\boldsymbol{x}_t^{\texttt{time}}),$$

where $\boldsymbol{x}_t^{\texttt{time}}$ is a vector of time-related features (e.g., time of day, day of the week) at time t. The regression function φ can be estimated from data, using any supervised learning (here, regression) technique. For example, if we are interested in predicting average future occupancy, φ can be obtained by minimizing the empirical mean square error, i.e., solving

$$\min_{\varphi} \frac{1}{T} \sum_{t=1}^{N} \left(z_{t+1} - N_t \times \varphi(\boldsymbol{x}_t^{\texttt{time}}) \right)^2, \tag{2}$$

while other loss functions can be used for quantile prediction (the pinball loss, see Sun et al. 2012). Recent methods apply hierarchical forecasting for predicting occupancy in several sub-units as well as in the entire ED, simultaneously and consistently (Bertani et al. 2025).

2.3.2. Aggregation of Patient-Level Predictions Alternatively, there is a growing interest in prediction of personalized survival probabilities (e.g., Gill et al. 2018, Bertsimas et al. 2021, Wang et al. 2022, Arora et al. 2023). In this setting, we leverage information about individual patients at each point in time, such as the severity of their condition or their current length of stay, which we denote $\boldsymbol{x}_{i,t}^{\texttt{patient}}$. We construct a model for the remaining occupancy z_{t+1} by using a two-step, bottom-up approach: First, we construct a model $\varphi(\boldsymbol{x}_{i,t}^{\texttt{patient}})$ to predict $\delta_{i,t+1}$. Then, we estimate z_{t+1} as

$$\hat{z}_{t+1}^{\texttt{bottom-up}} = \sum_{i:\delta_{i,t}=1} \varphi(\boldsymbol{x}_{i,t}^{\texttt{patient}}).$$

Note that $\boldsymbol{x}_{i,t}^{\texttt{patient}}$ may contain time-dependent covariates ($\boldsymbol{x}_t^{\texttt{time}}$) or information about other patients present in the ED (e.g., number of severe patients at time t). In particular, these features can capture interaction effects (e.g., slow-downs (Kc and Terwiesch 2009)) between patients.

An additional benefit from bottom-up approaches is their flexibility. Indeed, they can generate estimates for sub-units (e.g., patients in the observation area of the ED) or sub-categories of patients (e.g., high-acuity patients) by simply restricting the summation to a subset of $\{i : \delta_{i,t} = 1\}$. However, as we show analytically in Section 2.4, bottom-up approaches require careful calibration of the estimates of $\varphi(\boldsymbol{x}_{i,t}^{\texttt{patient}})$ to generate unbiased estimates, and calibration needs to be performed for each sub-group of interest.

2.3.3. Deep Set Models With the recent advances in deep learning, one could envision building a neural network model to predict z_{t+1} directly from the concatenation of all patient covariates, $\{\boldsymbol{x}_{i,t}^{\texttt{patient}}\}_{i:\delta_{i,t}=1}$. Since the prediction should be invariant to any permutation of the patient indexes, one can without loss of generality consider models of the form

$$\hat{z}_{t+1}^{\texttt{direct}} = \rho\left(\sum_i \boldsymbol{\phi}(\boldsymbol{x}_{i,t}^{\texttt{patient}})\right),$$

with arbitrary functions $\boldsymbol{\phi}$ and ρ (Zaheer et al. 2017, theorem 2). Estimators of these forms have been proposed in Zaheer et al. (2017) and referred to as deep set estimators. In particular, $\boldsymbol{\phi}(\boldsymbol{x})$ can be interpreted as an embedding of the patient information into a (potentially multi-dimensional) latent space.

This direct approach further generalizes the bottom-up approach described above, hence is potentially more powerful. For example, it can learn interactions between patients directly from data, while the bottom-up approach requires a priori feature engineering to do so. However, a deep set model may be harder to learn from data for at least two reasons: First, it requires to learn two functions, $\boldsymbol{\phi}$ and ρ, simultaneously, while the bottom-up approach fixes ρ to be the identity. Second, the encoding $\boldsymbol{\phi}$ is learned in a semi-supervised way, with the objective to minimize the prediction error of $\hat{z}_{t+1}^{\texttt{direct}}$, which is obtained after aggregating individual embeddings. On the contrary, the bottom-up approach considers a special case, where the embedding $\varphi(\boldsymbol{x}_{i,t}^{\texttt{patient}})$ is the solution of a patient-level prediction problem, and uses the patient-level labels $\delta_{i,t+1}$.

2.4. Bias when aggregating patient-level predictions and potential solutions

Unlike in the time-series and direct approaches, the predictive models used in the bottom-up approach are patient-level models specifically designed to predict outcomes at the patient level. These outcomes are closely related to, yet distinct from, the system-level quantity of interest, z_{t+1}. Several studies have documented the emergence of bias when aggregating patient-level predictions for occupancy forecasting (Bertsimas et al. 2021, Wang et al. 2022, King et al. 2022). We now expose the mechanisms that explain the occurrence of such bias, which are reminiscent of the inspection paradox in applied probability (Stein and Dattero 1985).

With our notations, the average bias of the prediction $\hat{z}_{t+1}^{\texttt{bottom-up}}$ can be expressed as

$$\frac{1}{T}\sum_{t=1}^{T}\left(z_{t+1} - \hat{z}_{t+1}^{\texttt{bottom-up}}\right) = \frac{1}{T}\sum_{(i,t)\in\mathcal{S}}\left(\delta_{i,t+1} - \varphi(\boldsymbol{x}_{i,t}^{\texttt{patient}})\right),$$

where $\mathcal{S} := \{(i,t) \in \{1,\ldots,N\} \times \{1,\ldots,T\} : \delta_{i,t} = 1\}$. So, achieving an average bias of zero can be expressed as

$$\frac{1}{|\mathcal{S}|}\sum_{(i,\tau)\in\mathcal{S}}\left(\delta_{i,t+1} - \varphi(\boldsymbol{x}_{i,t}^{\texttt{patient}})\right) = 0. \tag{3}$$

In other words, we see from (3) that one achieves unbiased predictions of the system-level quantity z_{t+1} if the patient-level predictions, $\varphi(\boldsymbol{x}_{i,t}^{\texttt{patient}})$, are unbiased estimates of $\delta_{i,t+1}$ over all $(i,t) \in \mathcal{S}$. Importantly, in the dataset $\mathcal{S}$, patient i contributes proportionately to their length of stay. Hence, a bias can emerge whenever there is a mismatch between the latter dataset and the one used for training the patient-level model φ.

Remark 1. Time-series approaches are not protected against bias issues either. For example, minimizing $\sum_t (z_{t+1} - N_t \times \varphi(\boldsymbol{x}_t^{\texttt{time}}))^2$ in (2) is not equivalent to minimizing $\sum_t (z_{t+1}/N_t - \varphi(\boldsymbol{x}_t^{\texttt{time}}))^2$. The latter objective corresponds to a weighted version of the former and leads, in particular, to biased predictions of z_{t+1}. Similarly, using loss functions that are less sensitive to outliers (e.g., mean absolute error) or sub/over-sampling—common practices when predicting skewed outcomes—could result in biased predictions. However, when considering time-series approaches, bias issues are more salient because the model is trained specifically to predict z_{t+1}. In the bottom-up approach, however, the patient-level model predicts an intermediate patient-level outcome that can be of interest to the practitioner in itself (and not only when aggregated across patients). Hence, bias in the downstream occupancy prediction task is more easily overlooked.

We now discuss the risk of bias of two specific patient-level outcomes identified in the literature because of their clinical relevance, which are also used for predicting future occupancy. Additionally, we propose solutions to mitigate these biases.

- **Remaining Length of stay** Length of stay (LoS) is widely considered as a proxy for quality of care (see Thomas et al. 1997, Brasel et al. 2007, for discussions on its relevance). Accordingly, hospitals are seeking models to predict patient LoS. For example, they can use EHR data to build a patient-level prediction, $\rho(\boldsymbol{x}_i^{\texttt{patient}})$, of $\lambda_i := \sum_t \delta_{i,t}$. In turn, at any point in time t, the same model can be used to estimate the remaining LoS for patient i, $\rho(\boldsymbol{x}_i^{\texttt{patient}}) - \sum_{s\leq t} \delta_{i,s}$, and construct a prediction of $\delta_{i,t+1}$ of the form $\varphi(\boldsymbol{x}_{i,t}^{\texttt{patient}}) = \mathbf{1}\left(\rho(\boldsymbol{x}_i^{\texttt{patient}}) - \sum_{s\leq t} \delta_{i,s} > \tau\right)$, with $\tau = 1$.

These models suffer from two main limitations. First, they estimate patient LoS upon arrival and do not update these predictions based on information revealed throughout the stay (e.g., time spent in the ED). Second, even when $\rho(\boldsymbol{x}_i^{\texttt{patient}})$ is an unbiased estimate of λ_i, $\rho(\boldsymbol{x}_i^{\texttt{patient}}) - \sum_{s\leq t} \delta_{i,s}$ is an over-estimate of the remaining LoS, as we formally state in Lemma 2 in Appendix A.

To address these limitations, we can build models $\phi(\boldsymbol{x}_{i,t}^{\texttt{patient}})$ that predict the remaining LoS, $\lambda_i - \sum_{s\leq t} \delta_{i,s}$, instead. These models can be used to predict $\delta_{i,t+1}$ via thresholding:

$$\varphi(\boldsymbol{x}_{i,t}^{\texttt{patient}}) = \mathbf{1}\left(\phi(\boldsymbol{x}_{i,t}^{\texttt{patient}}) > \tau\right).$$

In both cases, the resulting prediction on occupancy may not satisfy (3). While the threshold value $\tau = 1$ makes sense given the prediction task at hand (predicting remaining occupancy at time $t+1$), it does not guarantee unbiased occupancy predictions either. Instead, we propose two simple and effective calibration procedures to ensure condition (3). First, we can calibrate τ so as to achieve the same number of false positives as false negatives, hence satisfying (3) by design. Alternatively, we can use $\tau = 1$ followed by an affine transformation of the binary predictions, $\beta_0 + \beta_1 \times \sum_i \varphi(\boldsymbol{x}_{i,t}^{\texttt{patient}})$, where (β_0, β_1) are computed by solving a simple ordinary least square problem. By properties of ordinary least square regression, the resulting predictor is necessarily unbiased (in-sample)—see Lemma 4 in Appendix A.2.

- **Survival probabilities** Individual survival probabilities $\mathbb{P}(\delta_{i,t+1} = 1 | \delta_{i,t} = 1)$ —or entire survival curves $\{\mathbb{P}(\delta_{i,t+s} = 1 | \delta_{i,t} = 1),\, s \geq 1\}$— can also be used to monitor patients recovery and anticipate discharges. A desirable property for classification models is called calibration: A binary classification model is said to be *calibrated* when the predicted probabilities match the empirical frequency of event occurrence (Guo et al. 2017). We now show that using well-calibrated predicted probabilities is sufficient for satisfying (3). We denote $\rho(\boldsymbol{x}_{i,t}^{\texttt{patient}})$ the predicted probabilities (also referred to as classification scores). By partitioning the summation in (3) based on the value p that ρ can take, we have

$$\begin{aligned}\frac{1}{T}\sum_{(i,t)\in\mathcal{S}}\left(\delta_{i,t+1} - \rho(\boldsymbol{x}_{i,t}^{\texttt{patient}})\right) &= \frac{1}{T}\sum_{p}\sum_{(i,t)\in\mathcal{S}} \mathbf{1}\left(\rho(\boldsymbol{x}_{i,t}^{\texttt{patient}}) = p\right)\,\left(\delta_{i,t+1} - \rho(\boldsymbol{x}_{i,t}^{\texttt{patient}})\right)\\ &= \frac{1}{T}\sum_{p}\left[\sum_{(i,t)\in\mathcal{S}} \mathbf{1}\left(\rho(\boldsymbol{x}_{i,t}^{\texttt{patient}}) = p\right)\,\delta_{i,t+1} - p\sum_{(i,t)\in\mathcal{S}} \mathbf{1}\left(\rho(\boldsymbol{x}_{i,t}^{\texttt{patient}}) = p\right)\right],\end{aligned}$$

which equals zero if

$$p\sum_{(i,t)\in\mathcal{S}} \mathbf{1}\left(\rho(\boldsymbol{x}_{i,t}^{\texttt{patient}}) = p\right) = \sum_{(i,t)\in\mathcal{S}} \mathbf{1}\left(\rho(\boldsymbol{x}_{i,t}^{\texttt{patient}}) = p\right)\,\delta_{i,t+1},$$

for any value of p. This is precisely the definition of calibration for binary classification probabilities. Furthermore, a common measure of calibration is the Expected Calibration Error (ECE; see Guo et al. 2017, Naeini et al. 2015), which, in our context, is equal to

$$ECE(\mathcal{S}) := \frac{1}{|\mathcal{S}|}\sum_{p}\left|\sum_{(i,t)\in\mathcal{S}} \mathbf{1}\left(\rho(\boldsymbol{x}_{i,t}^{\texttt{patient}}) = p\right) - p\sum_{(i,t)\in\mathcal{S}} \mathbf{1}\left(\rho(\boldsymbol{x}_{i,t}^{\texttt{patient}}) = p\right)\right|.$$

Observe that the set of observations $\mathcal{S}$ is involved in the definition of the ECE. We refer to Online Supplement EC.1.5 for a more formal definition. With this definition, we can state the following result:

Lemma 1. *Assume that the patient-level predictions that are being aggregated, $\varphi(\boldsymbol{x}_{i,t}^{patient})$, are the predicted probabilities of a binary classification model. Then, for a sufficiently fine-grained discretization of the $[0,1]$ interval, we can bound the bias of $\hat{z}_{t+1}^{bottom\text{-}up}$ as follows*

$$\left|\frac{1}{T}\sum_{t=1}^{T} z_{t+1} - \hat{z}_{t+1}^{bottom\text{-}up}\right| \leq \frac{|\mathcal{S}|}{T} ECE(\mathcal{S}).$$

In other words, improved calibration necessarily leads to lower bias in the downstream prediction. Note that this result involves the ECE computed on the dataset $\mathcal{S}$, where each patient appears as many times as they stay in the hospital. This dataset might differ from the dataset used for training the binary classification model if, for example, sub/over-sampling techniques are used.

Alternatively, for binary outcomes, one can convert a continuous score into a binary prediction by using a user-specified threshold. In this case, as for the (remaining) LoS models, the condition (3) requires the threshold to be chosen so as to achieve the same number of false positives and false negatives. In practice, however, such thresholds are typically calibrated to achieve a target sensitivity or specificity level, or to maximize a given accuracy metric like the F- or Youden's J-score (Etu et al. 2022).

3. Time-Series Models Enriched with Side Information

Predicting future remaining occupancy is a task well-suited for time-series models. There is a large body of literature documenting the power of time-series models (see, e.g., Wargon et al. 2009, for a survey), especially when combined with external side information. In this section, we demonstrate the power of these approaches to our problem.

3.1. Data and models

We train linear (LR) and nonlinear (XGBoost and neural networks; XGB and NN respectively) models on calendar features. The calendar features include current occupancy, hour, day of the week, and month. To account for the cyclical nature of time, day of week, and month, we convert these features into angles between 0 and 2π and represent them with their cosine and sine values, leading to seven calendar features in total.

Many studies have investigated the value of external data in making hospital occupancy predictions. In their survey, Wargon et al. (2009) do not observe any significant improvement from adding weather-related variables. On the other hand, recent analyses find that adding internet search trends can improve model performance (Tideman et al. 2019, Trevino et al. 2022, Fan et al. 2022). Based on these findings, we did not include any meteorological data such as precipitation or ambient temperature but considered search trends variables from Naver Trends[3]. Naver Trends is the leading search engine in South Korea and gives us the relative search frequency of key query terms, from January 9, 2016 to December 31, 2019. We used similar search terms as Trevino et al. (2022), namely the name of our partner hospital, hospital, fever, cough, shortness of breath, numbness, weakness, stomach pain, chest pain, back pain, fine dust, and yellow dust. For simplicity, for predicting occupancy at a given time, we use search trends on the same day, leading to midly optimistic estimates of the predictive power of search trends.

[3] `www.naver.com`

Table 2 Out-of-sample results for predicting remaining occupancy using time-series methods.

Variables	Calendar			+ Search Trends			+ Patient Summary		
Model	LR	XGB	NN	LR	XGB	NN	LR	XGB	NN
MAE (patients)	5.57	5.39	5.54	6.35	5.02	6.18	4.78	3.84	**3.80**
MAPE	16.5%	12.5%	18.7%	16.5%	18.1%	20.8%	14.2%	12.5%	**12.4%**
R^2	0.686	0.618	0.695	0.612	0.739	0.616	0.764	0.848	**0.851**

Finally, we consider adding summary information about the patients currently in the ED, namely: the average time spent by the patients currently in the ED and the number of severe patients (KTAS level 1 or 2).

3.2. Empirical performance

We compare the performance of nine time-series approaches (three model types, three set of explanatory variables) on the test set. As expected given the limited number of predictive features, training these models can be done efficiently, within five minutes on a standard laptop. We measure predictive accuracy in terms of mean absolute error (MAE), mean absolute percentage error (MAPE), and the proportion of variance explained by the model (R^2).

The out-of-sample (i.e., on the test set) accuracy metrics for the time-series models are reported in Table 2. Based on these results, we make the following observations: First, we observe that, irrespective of the predictive variables used, using nonlinear models (especially, XGB) leads to a significant increase in predictive accuracy (10–20% improvement depending on the metric). Second, we observe little to no improvement from including search trends information, which is surprising compared to the results from the literature such as Trevino et al. (2022). This difference could be due to the fact that we are predicting remaining occupancy only (excluding new arrivals), that we performed less feature engineering and selection as Trevino et al. (2022), or by the fact that, unlike Google trends, Naver trends data are reported for the whole country and cannot be restricted to a specific metropolitan area. Finally, we observe that including—even a limited number of—information about patients currently in the ED leads to the most significant improvement. For the linear regression model (LR) for example, it reduces the MAPE from 16.5% to 14.2%, a 2.3 percentage points or 13.9% improvement. Similarly, XGB (resp. NN) with summary patient information achieves an 12.5% (resp. 12.4%) MAPE compared to 13.4% (resp. 18.7%) without. All in all, XGB or NN on calendar variables and summary patient information achieve the best performance, with an out-of-sample MAE of 3.80–3.84 patients and MAPE of 12.4–12.5%.

To further illustrate the importance of patient summary information, Figure 2 represents the feature importance the variables in the NN model, where importance is measured by the average information gain provided by each variable. While the most important feature is the current occupancy ('Current Occupancy'), we observe that the average time spent by all current ED patients ('Avg Time Spent in ED') is of comparable predictive power as day of the week or month.

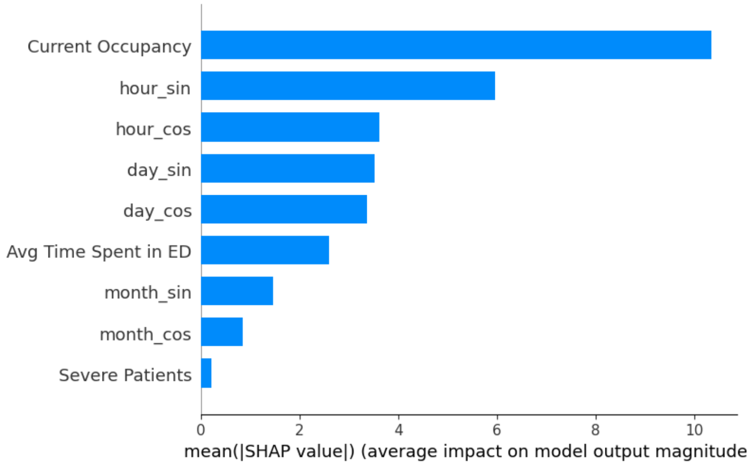


Figure 2 **Feature importance diagram for the time-series method using NN, calendar features and patient summary data.**

4. Aggregation of Patient-Level Predictions

In this section, we construct patient level models to predict patient remaining length of stay (rLoS), and aggregate them into an estimate for short-term ED occupancy. Our objectives are twofold: First, to empirically document the misalignment between the two predictive tasks (patient-level and system-level), which we described analytically in Section 2.4. Second, to demonstrate how simple calibration steps can be added and deployed in practice to effectively correct it.

Since this issue of misalignment and the resulting bias in remaining occupancy estimates can occur even in-sample (i.e., independently from considerations about over-fitting or nonstationarity), we compute and report all metrics (calibration and accuracy) on the training set in this section.

4.1. Possible patient-level models

Given our downstream objective to predict ED occupancy in the next four hours, we can define (and build models for) different patient-level outcomes to achieve this goal.

First, we can model the problem as a binary classification task: for every patient and every timepoint (e.g., every hour), we predict whether their rLoS exceeds four hours (rLoS > 4h). Alternatively, we can view each patient's rLoS as a continuous outcome and address this estimation task as a regression problem, or estimate the entire rLoS distribution via a survival model. These three approaches, classification/regression/survival analysis, correspond to the three broad families of models considered in the literature for patient (remaining) LoS estimation, which we review more extensively in Online Supplement EC.4. For the sake of brevity, we compare two modeling approaches here, classification and regression. Note that each model applies to a patient at a given point in time (every hour) and uses patient information (demographics and triage data described in

Section 2.2), ED-level information (occupancy and time-related features), and the time the patient has spent in the ED so far.

For classification models, we consider Random Forest (RF; Breiman 2001), gradient boosted trees (XGB; Chen and Guestrin 2016), and neural networks (NN). We measure patient-level accuracy using the area under the receiver operating characteristic curve (AUC) and calibration using the expected calibration error (ECE). For the regression approach, we predict the rLoS for each patient in the ED (in minutes) using a random forest regressor. Following practices from the literature, we consider predicting both the rLoS and its logarithm. We use mean absolute error (MAE), mean absolute percentage error (MAPE), and the proportion of variance explained by the model (R^2) as performance measures.

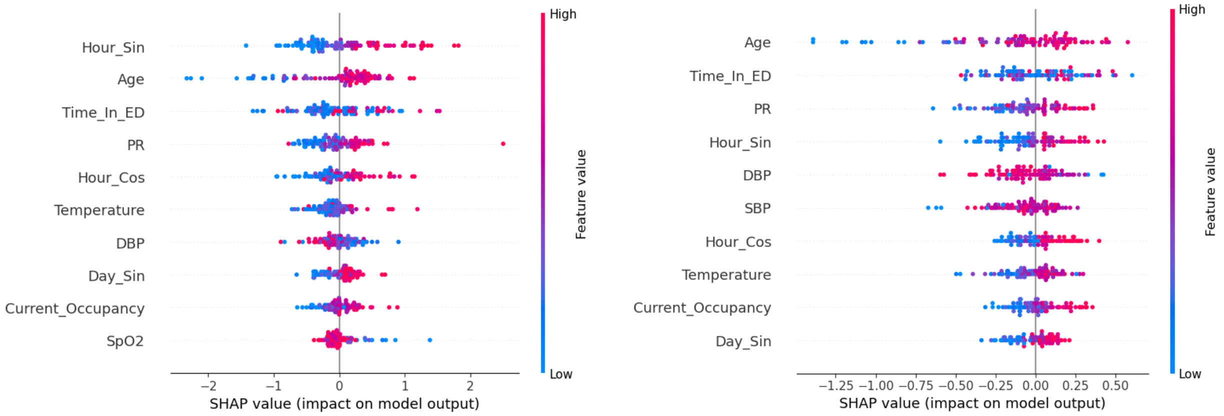
Compared with the time-series and the direct approaches, training bottom-up models is more time-consuming because of the larger number of observations—number of unique (patient, timepoint) pairs vs. T, the number timepoints—and the more precise prediction task. On our data, training either a classification or regression model and then aggregating their predictions to obtain an estimate for occupancy over three years of data requires around one hour.

Although both approaches (classification/regression) aim at capturing the same underlying outcome, they model a different part of the rLoS distribution. Conceptually, they group patients according to how close they are to the end of their visit, yet use different summary statistics of the rLoS distribution (one particular percentile and the mean, respectively). These differences in terms of what it means for patients to be similar lead to differences in the resulting models and how they should be interpreted. To illustrate these differences, Figure 3 displays the 10 most important features (SHAP plot) accordingly to a classification XGB model (left panel) and a regression XGB model (right panel). Although there are similarities in the most important variables (e.g., age, time spent in the ED until now, hour) and how they impact the predicted outcome, we also observe differences in the relative order of the features. Notably, the classification output seems to depend more on the hour of the day and the day of the week than its regression counterpart.

4.2. Calibrating classification models

As elicited in Section 2.4 (Lemma 1), binary classification models that are well calibrated (in the sense of calibrated probabilities) on the set of all combinations of (patient, timepoint) lead to unbiased prediction of future occupancy. Accordingly, we audit whether the predicted scores obtained by our raw classification models are calibrated. We apply a calibration correction step referred to as isotonic regression (see Online Supplement EC.1.4 for a description of isotonic regression) and evaluate its impact on patient-level classification and ED occupancy prediction accuracy.

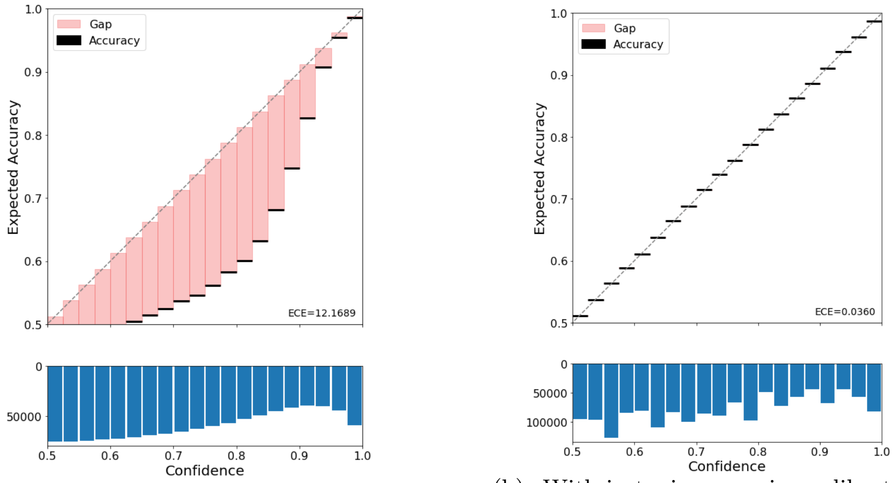
To visualize the impact of isotonic regression, Figure 4 displays the confidence vs. accuracy diagrams on the training set, for a NN classifier, before and after the isotonic regression step, and



(a) Classification

(b) Regression

Figure 3 Feature important plot (SHAP plot) for the XGBoost algorithm (XGB), for the classification (left panel) and regression (right panel) approaches.



(a) Without isotonic regression calibration

(b) With isotonic regression calibration

Figure 4 Calibration diagrams (on the training set) of the neural net classifier for predicting whether the remainig LoS of a given patient will exceeds four hours.

illustrates the significant gain in calibration (and corresponding improvement in ECE). The top part of Table 3 reports the AUC and ECE for the RF, XGB, and NN models, with and without isotonic regression, on the training set. As expected, we observe that isotonic regression significantly improves calibration, with no impact on the AUC (because scores are monotonously transformed). We also observe that the level of calibration for the raw scores varies significantly across models (0.94 for XGB vs. 12.2 for NN), emphasizing the need for data scientists to carefully audit the calibration of their models after training and to introduce a calibration step when needed.

In addition, the bottom part of Table 3 reports the accuracy of each model, when the outputs are aggregated across patients to estimate future occupancy. Our empirical results largely confirms

Table 3 Accuracy metrics (on the training set) of different patient-level classification models (rLoS > 4 h) and their corresponding accuracy on occupancy prediction.

Task		RF		XGB		NN	
		raw	+ isotonic	raw	+ isotonic	raw	+ isotonic
Patient-level	AUC	0.869	**0.869**	0.806	0.806	0.797	0.797
	ECE	7.33	0.09	0.94	0.00	12.2	0.036
Occupancy	MAE	3.35	**2.98**	3.13	3.11	3.15	3.13
	MAPE	11.9%	**9.89%**	10.7%	10.6%	11.0%	10.7%
	R^2	0.906	**0.940**	0.927	0.930	0.932	0.934
	Bias	0.128**	-0.054*	-0.003	-0.002	0.401**	0.007

*Notes: * = p-value* < 0.05*; ** = p-value* $< 10^{-3}$.

our theoretical findings: Non-calibrated predicted probabilities can result in biased occupancy predictions, even in-sample. On the other hand, applying an isotonic regression correction step, before the aggregation, not only improves the calibration of the predicted probabilities, but debiases occupancy predictions. It also moderately improves downstream accuracy. On the training data, the best performing model for the patient-level task is RF (AUC$= 0.869$ vs. 0.806 for XGB), which translates into the most accurate estimate for occupancy (in terms of MAE, MAPE, and R^2) *only when properly calibrated.*

Although miscalibrated probabilities or biased predictions can occur (and be fixed) in-sample, one should also be concerned about the out-of-sample behavior and, in particular, about the issue of overfitting for the isotonic regression model. For the sake of completeness, we report the out-of-sample analog of Table 3 in Appendix (Table B.1) and defer a more thorough discussion on out-of-sample performance and non-stationarity issues to Section 6.

4.3. Calibrating regression models

Similarly, regression models for predicting patient rLoS need to be carefully aggregated. In the top part of Table 4, we report the accuracy of two models (i.e., for predicting rLoS or its logarithm, with RF), on the training set. Intuitively, errors of the regression model will translate into false positive/negative rates of the binary prediction. As showed in Section 2.4, calibration in this setting aims at equaling these two types of error in order to get unbiased estimates of occupancy. We evaluate three aggregation strategies, as proposed in Section 2.4:

- (None) We convert individual predictions on remaining LoS (in minutes), $\hat{\ell}_{i,t}$, into binary predictions by testing whether $\hat{\ell}_{i,t}$ is greater than or equal to 240 minutes. We then sum these binary predictions over all patients present in the ED: $\sum_i \mathbf{1}(\hat{\ell}_{i,t} \geq 240)$.
- (LR correction) We apply the naive aggregation strategy described previously but instead return $\beta_0 + \beta_1 \times \sum_i \mathbf{1}(\hat{\ell}_{i,t} \geq 240)$, where (β_0, β_1) are the coefficients of a simple linear regression model that predicts future occupancy based on the naive aggregate estimate.

Table 4 Accuracy metrics (on the training set) of different patient-level random forest regression models and their corresponding accuracy on occupancy prediction, with/without calibration strategies. For log(rLOS) models, patient-level accuracy on rLOS (i.e., after taking the exponential of the predictions) is reported.

	Outcome	rLoS			log(rLoS)		
	Calibration	None	LR correction	Threshold	None	LR correction	Threshold
Patient	MAE	73.0	-	-	143	-	-
	MAPE	20.3%	-	-	23.1%	-	-
	R^2	0.98	-	-	0.81	-	-
Occupancy	MAE	4.19	2.55	2.21	1.54	1.44	1.28
	MAPE	13.1%	8.04%	6.95%	5.13%	4.83%	4.11%
	R^2	0.895	0.953	0.964	0.983	0.986	0.988
	Bias	4.13**	-0.0006	0.225**	-0.902**	-0.0006	0.169**

*Notes: * = p-value* < 0.05*; ** = p-value* $< 10^{-3}$.

- (Threshold) We apply a threshold-based rule to individual predictions on rLoS and sum these binary predictions together, $\sum_i \mathbf{1}(\hat{\ell}_{i,t} \geq \tau)$, but do not fix the threshold, τ, at 240 minutes. Instead, we calibrate it so that the binary output makes as many false negative mistakes as false positive ones when classifiying $\delta_{i,t+1}$. We find that using a threshold of $\tau = 285$ minutes for rLoS and $\tau = 225$ minutes for log(rLoS) equalize the number of false negatives and false positives on the training set.

The respective accuracy of each aggregation strategy is reported in the bottom half of Table 4. We observe that using a log-transformation leads to more accurate predictions (in terms of absolute and relative errors) of the remaining occupancy, for all aggregation strategies. When no calibration is used (None), we observe that the estimates of future occupancy with both models are significantly biased: The rLoS model underestimates the actual occupancy by 4.1 patients on average, while the log(rLoS) ones overestimates it by 0.9 patients. On this regard, both calibration strategies are effective in reducing this bias, with the LR correction being more effective. We believe this is due to the discrete nature of the threshold in the threshold strategy, which cannot always perfectly equalize false negatives and positives. Second, we observe that these calibration steps also significantly improve the accuracy obtained with each model, with the model for rLoS (unsurprisingly) benefiting the most. This improvement, however, is not sufficient to outperform the predictions made from the model for log(rLoS).

5. Direct Prediction using Deep Set Models

At any point in time, the ensemble of patients present in the ED can be considered as a set, where each patient is characterized by patient-specific covariates, as well as attributes that are shared across the cohort, such as the time of day and current occupancy. Deep set networks (Zaheer et al.

2017) constitute a particular type of neural network architecture designed to be applied to set-type inputs of potentially varying size. In this section, we describe how to implement the deep set architecture in our context and analyze its performance.

5.1. Architecture and implementation

As described in Section 2.3, the prediction returned by a deep set architecture is of the form

$$\rho\left(\sum_i \boldsymbol{\phi}(\boldsymbol{x}_{i,t}^{\texttt{patient}})\right).$$

The function $\boldsymbol{\phi}$ transforms each patient-information $\boldsymbol{x}_{i,t}^{\texttt{patient}}$ into a d-dimensional latent space. In particular, this transformation is the same for all patients. Then, these latent representations are summed together and another nonlinear function ρ is applied. In particular, under some assumptions, Zaheer et al. (2017) show that any function that supports input sets of varying cardinality and that is invariant to the indexing of the elements in that set can be represented in that form.

Formally, we train $\boldsymbol{\phi}$ and ρ so as to predict z^{t+1} based on $\{\boldsymbol{x}_{i,t}^{\texttt{patient}}, i = 1, \dots I\}$, where I is an upper-bound on the total number of patients that can be present in the ED at any point in time ($\max_t N_t \le I$) and $\boldsymbol{x}_{i,t}^{\texttt{patient}} \in \mathbb{R}^p$ is the vector of patient covariates. To address the issue of having a different number of patients at different point in time, we set $\boldsymbol{x}_{i,t}^{\texttt{patient}} = \mathbf{0}$ for $i = N_t + 1, \dots, I$.

In terms of architecture, we consider a function $\boldsymbol{\phi}$ represented as a neural network with one hidden layer with 64 units and one output layer with d units, hence comprised of $(p+1) \times 64 + (64+1) \times d$ tunable parameters. The function ρ takes as input $\sum_{\boldsymbol{x}} \boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^d$ into a hidden layer of size d and one output layer of size 1, leading to $(d+1) \times d + (d+1)$ parameters. We acknowledge that further refinement of the network architecture could lead to increased performance. In our experiments, we restricted ourselves this structure with 64 hidden nodes to avoid compression in the first layer (as $p = 18 \le 64$) since wider networks tend to exhibit more feature learning (Yang and Hu 2020). We denote these architectures as p–64–d and d–d–1, respectively. In the experiments of the following section, we focus on an architecture with $d = 64$, which results in $(18+1) \times 64 + (64+1) \times 64 = 1,216 + 4,160 = 5,376$ parameters for $\boldsymbol{\phi}$ and $(64+1) \times 64 + (64+1) = 4,225$ parameters for ρ. We also investigate the impact of the latent space dimension d and compare with architectures that mimic the bottom-up approach from Section 4. We train all models for 1,000 epochs with early stopping based on MAE on a holdout validation set consisting of 15% of the 2016-2018 training data.

5.2. Numerical performance

Table 5 reports the out-of-sample performance (in-sample performance is reported in Table B.3) of a deep set estimator with the architecture $\boldsymbol{\phi} : p$–64–d and $\rho : d$–d–1 for $d = 64$.

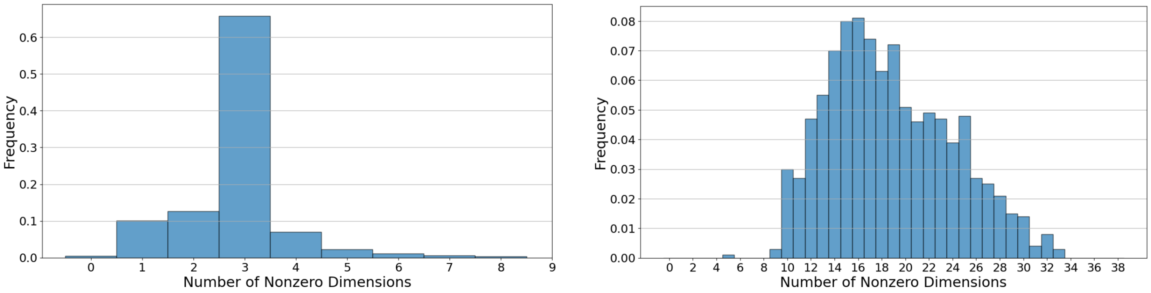
Table 5 **Out-of-sample results for predicting remaining occupancy using direct methods neural network with latent dimension of 1 and 64.**

Latent dimension	$d = 64$	$d = 1$	$d = 1$
$\boldsymbol{\phi}$	p–64–64	p–64–64–1	p–64–64–64–1
ρ	64–64–1	1–1	1–1
# Parameters	9,601	5,443	9,603
# Layers	4	4	5
MAE (beds)	**3.87**	3.91	4.05
MAPE	**48.2%**	49.0%	49.2%
R^2	**0.841**	0.839	0.823

As previously observed, the bottom-up approach from Section 4 can be seen as a special case of deep set estimators with additional structure. In particular, the bottom-up approach leads to architectures with $d = 1$ ($\boldsymbol{\phi}(\boldsymbol{x})$ corresponds a patient-level probability) and ρ is the identity. In addition, in a bottom-up approach, $\boldsymbol{\phi}$ and ρ are trained sequentially, with $\boldsymbol{\phi}$ learned to solve a patient-level task (i.e., predict the patient-level labels $\delta_{i,t+1}$), while, in the deep set approach, they are learned simultaneously for predicting the system-level quantity z_{t+1}. To illustrate the difference between the bottom-up and the direct approaches, we also consider two architectures that mimic the bottom-up ones as closely as possible, i.e., with $d = 1$ and $\rho: 1$–1. To allow for a fair comparison with the architecture with $d = 64$, we consider both $\boldsymbol{\phi}: p$–64–64–1 (same total number of layers) and $\boldsymbol{\phi}: p$–64–64–64–1 (comparable number of parameters).

We observe in Table 5 that our architecture with $d > 1$ improves out-of-sample accuracy compared with the two architectures with $d = 1$, highlighting the benefit from a richer patient embedding. To better appreciate the richness of the information captured by $\boldsymbol{\phi}(\boldsymbol{x})$, Figure 5 represents the distribution in the number of nonzero coordinates of $\boldsymbol{\phi}(\boldsymbol{x})$ (Figure 5(a)) and of the input to ρ, $\sum_{\boldsymbol{x}} \boldsymbol{\phi}(\boldsymbol{x})$ (Figure 5(b)). We observe that most patients are described using about 1–5 latent features (note that the embedding of $\boldsymbol{\phi}(\mathbf{0})$ also contributes to the proportion of patients with an embedding of size 3), leading to approximately 10–30 variables to describe the state of the ED and that are passed to ρ. This observation does not imply that the same 10–30 coordinates of $\boldsymbol{\phi}(\boldsymbol{x})$ are passed to ρ. A more fine-grained analysis (Figure B.2) shows that around 45 coordinates of the embedding are ever used. To confirm these findings, we also conduct supplementary experiments where we observe a noticeable gain in accuracy for $d \geq 8$ (see Figure B.3).

One can legitimately wonder to what extent the embedding $\boldsymbol{\phi}$ learned by the deep set estimator is specific to the task of occupancy prediction or whether it can be used as a generic embedding of patient state. To do so, we re-use the three embeddings from Table 5 for the patient-level classification task of predicting $\delta_{i,t+1}$. Specifically, when $d = 64$, we train a simple logistic regression model to predict $\delta_{i,t+1}$ as a function of $\boldsymbol{\phi}(\boldsymbol{x})$, and when $d = 1$, we use $\boldsymbol{\phi}(\boldsymbol{x})$ as a classification score



(a) Sparsity of $\boldsymbol{\phi}(\boldsymbol{x})$

(b) Sparsity of $\sum_{\boldsymbol{x}} \boldsymbol{\phi}(\boldsymbol{x})$

Figure 5 Empirical distribution (histogram) of the number of nonzero coordinates of $\phi(\boldsymbol{x})$ (left panel) and $\sum_{\boldsymbol{x}} \phi(\boldsymbol{x})$ (right panel). Results are obtained by sampling 1,000 random time points t.

Table 6 AUC (in- and out-of-sample) obtained on the classification task of predicting $\delta_{i,t+1}$ using the embedding ϕ obtained from the deep set estimator for remaining occupancy.

Latent dimension	$d = 64$	$d = 1$	$d = 1$
$\boldsymbol{\phi}$	p–64–64	p–64–64–1	p–64–64–64–1
In-Sample	0.697	0.752	0.753
Out-of-Sample	0.633	0.687	0.686

directly. Table 6 reports the AUC for each of these approaches. While the performance levels are acceptable (0.70–0.75 in-sample, 0.63–0.69 out-of-sample), there are significantly lower than that of dedicated classification models (Section 4.2). Unsurprisingly, one-dimensional embeddings perform best, suggesting that, when constrained to $d = 1$, $\boldsymbol{\phi}$ tries to recover $\delta_{i,t+1}$. Nonetheless, models that have access to $\delta_{i,t+1}$ explicitly during training demonstrate a significant edge.

Overall, the predictive performance of the deep set approach reported in Table 5 is on par with that of bottom-up models from Section 4.2 (we defer a formal comparison of out-of-sample performance to Section 6.1), except for a relatively high MAPE. Since MAPE measures relative errors, it is sensitive to errors made when the true value (i.e., future occupancy) is small. Indeed, we observe in Figure B.5 that the deep set model performs poorly in some (rare) cases with low future occupancy (≤ 20 beds), resulting in a large impact on the overall MAPE.

6. Comparison of Predictive Accuracy in Non-Stationary Environment

We can now compare the performance of time-series models with that of bottom-up and direct approaches. To be practically relevant, we evaluate and compare their performance out-of-sample, on data from 2019, in Section 6.1.

However, as suggested by Figure 1(b), our environment experiences distributional shifts over time (like many real-world environments). Distribution shift, which generally refers to changes in data distribution between training and testing phases, is a serious hurdle in the adoption of machine learning systems in healthcare (Challen et al. 2019, Riley 2019). For example, Nestor et al. (2019) observe a 0.3 (resp. 0.1) drop in AUC for mortality (resp. length of stay) prediction due to changes in the record-keeping practices at a hospital over time. We discuss how such shifts can be detected and accounted for in Section 6.2.

6.1. Comparison of out-of-sample performance

To capture the inherent nonstationarity of our prediction task and provide a realistic comparison, we mimic a setting where a model has been trained on past data (2016-2018) and is being deployed in production in 2019. For each approach, we restrict our attention to the best performing models, as well as neural network models to allow for a better apples-to-apples comparison across approaches[4]. All in all, we consider two time-series models (XGB and NN) with calendar variables and patient summary information; two classification models for patient rLoS (XGB and NN), calibrated with isotonic regression and then aggregated[5]; one direct model consisting of a deep set model with a latent space of dimension 64.

Table 7 reports the accuracy of each model on data from 2019. First, we should acknowledge the remarkable performance of the time-series models, which, despite the simplicity of the variables they use, explain 85% of the variance in remaining occupancy and are the two best performing methods in terms of MAE and MAPE. We should emphasize that this performance is largely due to the presence of summary patient information, in particular the average LoS of current ED patients. Second, the performance of the bottom-up and the direct approach are comparable, with an edge for bottom-up approaches, especially in terms of MAPE.

Regarding the bottom-up approaches, we observe that calibration of the intermediate probabilities is beneficial for downstream occupancy prediction, although the benefits are less acute out-of-sample than in-sample (Table 4). Despite our best efforts in Section 4, the resulting predictions on remaining occupancy remain biased out-of-sample. This is due to the fact that the isotonic regression model is fitted on the training data. As any model learned from data, it can suffer from over-fitting or become less relevant with time when applied in a nonstationary environment, as we discuss in the following section.

6.2. Detecting and mitigating distribution shifts

Figure 1(b) indicates a shift in the distribution of our outcome of interest (here, the probability of a patient currently in the ED staying at least four hours). This distribution shift could be due

[4] For the time-series and bottom-up approaches, we consider a NN architecture with number of parameters comparable with the NN for the direct approach, namely p–64–64–64–1.

[5] We report the performance of the bottom-up approach with the two RF models, with a classification and a regression patient-level task respectively, in Table B.2 in Appendix.

Table 7 Out-of-sample accuracy for predicting remaining occupancy with a time-series approach (using calendar and patient summary features), a bottom-up approach (with an intermediate patient-level classification task), and a direct approach (with a latent space of size $d = 64$). Models are trained on 2016-2018 data and evaluated on data from 2019.

Approach	Time-Series		Bottom-Up				Direct
			Raw		+Calibration		
Model	XGB	NN	XGB	NN	XGB	NN	
MAE (patients)	3.84	**3.80**	3.82	4.25	3.85	4.09	3.87
MAPE	**12.5%**	12.4%	12.7%	13.9%	12.7%	13.3%	48.2%
R^2	0.848	**0.851**	0.847	0.804	0.850	0.816	0.841
Bias	**+0.62**	+0.77	+2.05	+2.19	+2.02	+1.26	+0.78

to nonstationarity in the patient population (covariate shifts) or in the relationship between the patient covariates and their (remaining) LoS. In the presence of nonstationarity, the performance of any predictive models is likely to deteriorate over time. In this section, we briefly present how to properly detect distribution shifts, especially when using a bottom-up approach, and evaluate the benefits of simple retraining strategies.

A standard approach to detect covariate shift is to compare the distribution the covariates used by the models, on the training and test sets. For the patient-level models, many explanatory variables are patient-specific and do not change over time, such as demographic information (age, sex) or vital signs at triage. Consequently, we can test whether their mean (t-test) or their distribution (Kolmogorov-Smirnov test) is the same for patients in the training set as for patients in the test set. However, in our case, this approach is flawed. Indeed, as elicited in Section 2.4, when used to predict future occupancy, the patient-level models will be applied to a dataset where each patient contributes proportionately to their LoS ($\mathcal{S}$ in the notations of Section 2.3). Hence, the statistical tests need to be conducted on a dataset where patients are not equally weighted but weighted proportionately to their LoS.

This distinction can completely overturn the conclusion of the statistical tests. Theoretically, we show (Lemma EC.1 in Online Supplement EC.2) that a t-test can identify a statistically significant change in mean when observations are equally weighted but no significant change when weighted according to LoS, and vice-versa. Empirically, we conduct t- and KS-tests to detect distribution change in our patient-level variables. For two patient-level variables, namely temperature at triage and sex indicator, we cannot reject the null hypothesis for at least one of the tests when patients are weighted equally (as is mostly done for patient-level information). However, when weighting patients by their LoS (which, again, is the relevant weighting for our occupancy prediction task), we detect statistically significant changes in the distributions of both variables, as reported in Table 8. For the remaining six variables, we observe a significant distribution change (p-value

Table 8 Results from t- and Kolomogorov-Smirnov (KS) tests to detect nonstationarity in the distribution of Temperature and Sex F between the training and testing data, depending how the weights that are applied to each patient.

Variable	Temperature		Female	
Weight per patient	Equal	LoS	Equal	LoS
Average on 2016-2018	37.04	37.18	0.51	0.47
Average on 2019	37.04	37.13	0.51	0.49
t-test p-value	1.00	$< 10^{-4}$	0.71	$< 10^{-4}$
KS-test p-value	$< 10^{-4}$	$< 10^{-4}$	1.00	$< 10^{-4}$

$\leq 5 \cdot 10^{-3}$) irrespective of the weight applied to each observation (see Tables EC.3(a)–(b) in Online Supplement).

To mitigate the impact of nonstationarity on the performance of the model deployed in production, retraining the models using the most recent data can be an effective strategy (Gama et al. 2014). We now adapt this solution to our three approaches. For the time-series approach, we retrain the models every month of 2019, including the past month data in the training data set. While viable for simple models that require a limited number of observations and variables, retraining can be prohibitively time-consuming for more sophisticated patient-level models like the ones we constructed in Section 4. Instead, we propose to apply this retraining strategy to the calibration step only. Specifically, we use a patient-level model trained using 2016-2018 data and we retrain the calibration step of the model every month, using the data from the past month. Our proposal echoes the observations made by Davis et al. (2017)—although in a different healthcare context, namely risk of hospital-acquired acute kidney injury—that classification models sometimes decalibrate over time, without observable impact on pure accuracy. For the direct approach, we adopt a similar strategy in the sense that we do not retrain the embedding $\boldsymbol{\phi}$ but only the aggregation function ρ. Instead of a complete retraining from scratch, however, we adopt a *fine-tuning* strategy, inspired by recent evidence from the literature (e.g., Lee et al. 2022), where we use data from the past month to fine-tune the values parameters of the current model.

Table 9 reports the accuracy of each approach with monthly retraining, on data from 2019. First, we observe that a rolling calibration strategy is effective at mitigating the impact of nonstationarity on patient-level models, as it noticeably improves their accuracy (around -0.4 reduction in MAE for bottom-up approaches, -0.15 for the direct approach) without requiring to retrain the entire model. Second, we observe that the bottom-up and direct approaches benefit the most from retraining compared with the time-series model. We suspect the relative robustness of the time-series model could be explained by its auto-regressive nature and the fact that the predictions naturally adapts to past realizations. Finally, we observe that time-series models are no longer the most accurate methods once retraining strategies are considered. Deep set estimators with a monthly fine-tuned ρ

Table 9 Out-of-sample accuracy for predicting remaining occupancy for the same models as Table 7 with a retraining strategy. Models are trained on 2016–2018 data and evaluated on data from 2019, with monthly retraining.

Approach	Time-Series		Bottom-Up		Direct
Model	XGB	NN	XGB	NN	NN
MAE	3.80	3.79	**3.48**	3.66	3.72
MAPE	12.4%	12.4%	**11.3%**	11.9%	46.9%
R^2	0.850	0.855	**0.862**	0.840	0.854
Bias	+0.56	+0.84	+0.07	–0.08	**+0.02**

are slightly more accurate in terms of MAE (-0.07). Overall, bottom-up approaches with predicted patient-level probabilities of staying four additional hours lead to the best performance. With XGB as a patient-level classifier, it achieves an MAE, MAPE, and R^2 of 3.48, 11.3%, and 0.862 respectively. In particular, it outperforms the time-series and direct approach by 0.25–0.3 in MAE.

7. Extensions to Multiple Prediction Horizons and Emergency Departments

Our main analysis applies to the task of predicting ED occupancy in 4 hours at our partner hospital. In this section, we evaluate the robustness of our findings across different time horizons and healthcare settings. In Section 7.1, on the same hospital data as our main analysis, we evaluate the impact of varying the prediction horizon from 1 to 12 hours. In Section 7.2, we replicate our analysis to 10 Korean hospitals and verify our findings on EDs of varying sizes and characteristics.

7.1. Performance Across Multiple Prediction Horizons

Instead of focusing on 4-hour predictions only, we now evaluate the performance of our three modeling approaches—time-series, bottom-up, and direct—for prediction horizons ranging from 1 hour to 12 hours. For each approach, we use neural networks models and replicate the out-of-sample analysis of Section 6.2. In particular, accuracy metrics are reported for one year of implementation with monthly retraining/recalibration.

Figure 6 shows the trend in MAE for each modeling approach (with a NN model), as the prediction horizon increases. A more comprehensive comparison of additional performance metrics is provided in Online Supplement EC.5.1. Overall, we observe that our findings are robust to the prediction horizon: The bottom-up and the direct (deep set) approaches achieve the lowest MAE consistently across all prediction horizons. The time-series approach, while performing adequately at shorter prediction horizons, has the highest MAE at all prediction horizons. Overall, we observe that the relative benefit of using a bottom-up (resp. direct) approach over a time-series model increases with the prediction horizon, leading to a decrease in MAE of 3.3% (resp. 0.8%) for 1-hour ahead predictions to 11.8% (resp. 19.1%) for 12-hour ahead predictions. Between the bottom-up

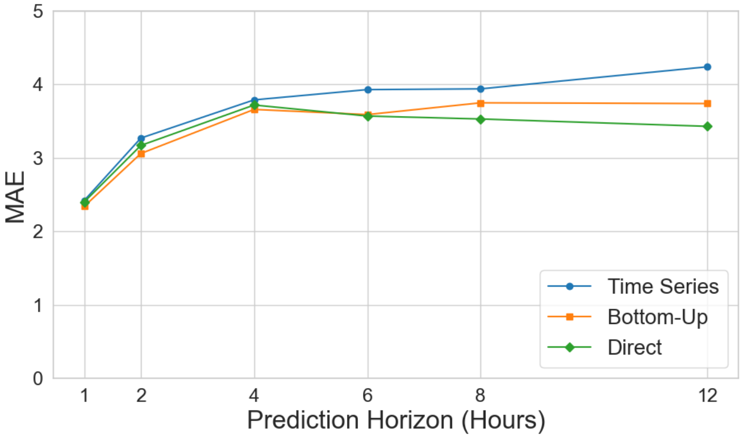


Figure 6 Out-of-sample MAE of the three modeling approaches as a function of the prediction horizon.

and direct approaches, we observe that the direct is superior in terms of MAE for longer prediction horizons. However, as already observed in Sections 5.2 and 6.2, direct models can perform poorly on hours with very low occupancy, leading to poor performance in terms of MAPE (Figure EC.1).

7.2. Generalization to Multiple Emergency Departments

We extend our analysis to multiple medical centers. For this analysis, we use data from the National Emergency Department Information System (NEDIS) in South Korea, a nationwide, real-time database that aggregates ED data from over 400 hospitals across the country (Yoo et al. 2023). The database includes detailed information on patient demographics, triage scores, chief complaints, diagnoses, treatment procedures, ED arrival and discharge times, disposition outcomes, and mode of arrival, covering the 2018–2022 period. Unfortunately, the database does not include features that vary along the visit.

We focus our analysis on EDs located in the Seoul metropolitan area, which includes a total of 50 EDs. Seoul contains a high concentration of high-volume, tertiary, and secondary hospitals, making it an ideal setting to evaluate the performance of occupancy prediction models in a crowded urban environment. Among these, we select the 10 EDs with the highest number of visits in 2022, as these sites are often overcrowded and thus are likely to benefit the most from improved occupancy prediction, but excluding our partner hospital, the original study site, to ensure independent external validation of our findings.

Table 10 presents some summary statistics for these EDs. In our data, we observe that larger EDs tend to have longer ED visits, a trend that is well documented in the literature. This is commonly attributed to higher patient volumes, greater case complexity (see Table EC.2 for average LOS by KTAS levels), and more frequent crowding and boarding in larger or academic centers (Karaca

Table 10 **Summary statistics for each of the 10 EDs in our database (sorted by hourly occupancy).**

	Daily Avg	Avg Hourly	Avg LOS	% LOS	KTAS (%)				
ED	Visits	Occupancy	(min)	> 4 hrs	1	2	3	4	5
1	108.14	9.34	116.90	11.63%	0.81	3.75	28.07	35.94	31.45
2	117.54	11.58	155.47	16.98%	0.91	6.80	51.98	31.69	8.62
3	127.61	15.73	177.51	20.72%	0.88	3.69	32.83	54.99	7.61
4	129.77	17.05	189.42	23.67%	0.60	4.31	35.82	33.95	25.30
5	130.98	17.08	189.22	25.48%	0.76	7.55	50.81	33.40	7.48
6	144.19	26.76	269.08	36.44%	0.66	3.77	53.85	25.83	15.88
7	161.18	39.78	357.50	36.28%	0.85	5.36	46.42	40.48	6.85
8	168.42	45.55	391.12	49.08%	2.00	10.27	50.48	30.74	6.51
9	230.76	63.78	399.62	48.51%	1.40	8.11	27.82	51.90	10.77
10	288.31	83.43	417.30	49.71%	2.02	8.75	56.33	29.39	3.52
Average	153.66	33.01	302.66	36.24%	1.24	6.74	44.14	37.10	10.78

et al. 2012). Teaching hospitals, which are typically larger, also tend to have longer lengths of stay due to the involvement of trainees and higher rates of diagnostic testing and consultations (Karaca et al. 2012, Riguzzi et al. 2014). Although we restricted our attention to the top-10 EDs in terms of 2022 annual volume, we observe that the hourly occupancy ranges from 10 to 80 patients on average, providing a good coverage of ED sizes. The models were trained using data from 2018, 2019, and 2021, while data from 2022 was reserved for testing. Compared to our main analysis in Section 6, the data for this extension covers both pre- and post-COVID periods. Neural networks were again used for all hospitals and approaches.

Figure 7 shows the out-of-sample MAE for the 10 EDs and a 4-hour prediction horizon. Additional performance metrics, including MAPE, R^2, and bias, are reported in Figure EC.2. We observe that the time-series approach shows the highest MAE, across all 10 EDs. In comparison, the bottom-up or direct approaches do provide a gain in accuracy. However, the benefit appears marginal for the smaller EDs (0–10% reduction in MAE for ED1–5) and much more acute for the larger ones (15–35% MAE reduction for ED6–9). Comparing bottom-up and direct, both approaches perform similarly on small EDs but their performance diverges on larger ones.

Our results indicate that the time-series approach, while commonly used, appears to be less suited for larger EDs. The bottom-up and direct approaches yield comparable performance, with no clear winner. These findings highlight that model performance varies across institutions, suggesting the need for customized model calibration rather than a one-size-fits-all approach.

8. Conclusion

With the increasing availability of patient-level data and the commoditization of data analytics tools, hospitals are eager to build prediction models that leverage the rich patient-level data they

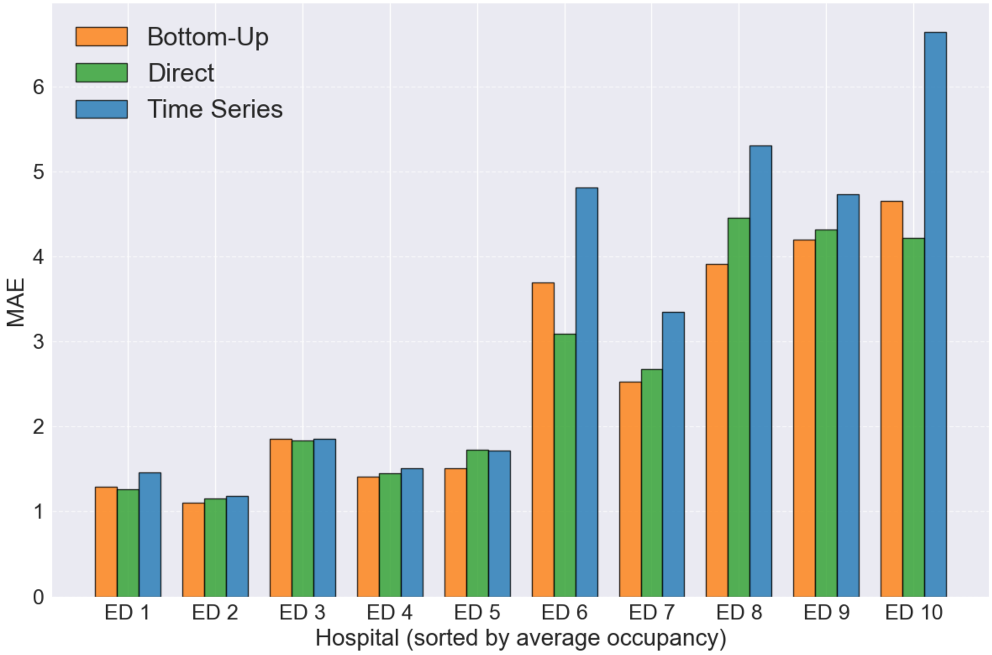


Figure 7 **Out-of-sample MAE for 4-hour occupancy prediction, across our three forecasting approaches and our 10 different EDs.**

collect. Our study presents and compares models that can benefit from patient-level information to predict ED occupancy.

First, we investigate the relevance of time-series approaches. We find that including summary information about patients currently in the ED is extremely valuable for future occupancy prediction—much more than external data such as search trends, in our case study— leading to a 15–30% reduction in MAE. Among other advantages, these approaches are simple (around 10 predictive variables) and relatively robust to distribution shifts compared with other models.

Second, we consider bottom-up pipelines that first predict a patient-level outcome (such as the probability of staying at least four hours) and then aggregate these predictions across all patients present in the ED. Because they provide two levels of prediction, at a patient and ED level, these approaches are more interpretable from a clinical perspective. Eventually, we find that they lead to the most accurate occupancy predictions, improving MAE by 0.3 patients (or 35%) compared to time-series models, at the expense of a more expensive training effort. To achieve these accuracy gains, however, a key finding (analytical and empirical) from our analysis is that the patient-level predictions need to be carefully calibrated, and re-calibrated over time to achieve those gains.

Finally, deep sets are a promising family of models that can predict occupancy directly from the concatenation of all patient-level information available from the ED. In our case study, they achieve an intermediate performance between time-series and bottom-up approach, provided that they are dynamically fine-tuned over time (every month in our experiments). However, we observe that these models perform poorly on low-volume days (leading to very high overall MAPE), suggesting they might not be well suited for smaller institutions.

Overall, our findings illustrate the benefits of using patient-level data for predicting ED occupancy. However, it also underscores the importance of attention to model selection, calibration, and fine-tuning processes. We hope this study offers a practical guide for healthcare analytics teams aiming to leverage patient-level data for enhancing operational efficiency of healthcare systems.

Beyond healthcare, in many customer service system (e.g., call centers or e-commerce websites), data is now available at a granular level (e.g, at an interaction or transaction level) and enables predictions of individual outcomes such as probability of completion or purchase. Yet, for some strategic or tactical decisions, predictions on aggregate quantities (e.g., number of available operators or total demand) might be more relevant to managers. In this context, we believe our analysis and insights could interest researchers and practitioners in service systems more broadly, about the opportunities and challenges posed by granular-level information for system-level forecasting.

References

Arora S, W Taylor J, Mak HY (2023) Probabilistic forecasting of patient waiting times in an emergency department. *Manufacturing & Service Operations Management* .

Awad A, Bader-El-Den MB, McNicholas J (2017) Patient length of stay and mortality prediction: A survey. *Health Services Management Research* 30:105 – 120.

Bacchi S, Tan Y, Oakden-Rayner L, Jannes J, Kleinig T, Koblar S (2022) Machine learning in the prediction of medical inpatient length of stay. *Internal Medicine Journal* 52(2):176–185.

Bertani N, Jensen ST, Satopää VA (2025) Joint bottom-up method for probabilistic forecasting of hierarchical time series. *Operations Research* .

Bertsimas D, Pauphilet J, Stevens J, Tandon M (2021) Predicting inpatient flow at a major hospital using interpretable analytics. *Manufacturing & Service Operations Management* .

Borges D, Nascimento MC (2022) COVID-19 icu demand forecasting: A two-stage prophet-lstm approach. *Applied Soft Computing* 125:109181, ISSN 1568-4946.

Brasel KJ, Lim HJ, Nirula R, Weigelt JA (2007) Length of stay: an appropriate quality measure? *Archives of Surgery* 142(5):461–466.

Breiman L (2001) Random forests. *Machine Learning* 45:5–32.

Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K (2019) Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety* 28(3):231–237, ISSN 2044-5415.

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.

Cheng Q, Argon NT, Evans CS, Liu Y, Platts-Mills TF, Ziya S (2021) Forecasting emergency department hourly occupancy using time series analysis. *The American Journal of Emergency Medicine* 48:177–182.

Davis S, Lasko T, Chen G, Siew E, Matheny M (2017) Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association* 24.

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), ISSN 2076-3417.

Etu EE, Monplaisir L, Arslanturk S, Masoud S, Aguwa C, Markevych I, Miller J (2022) Prediction of length of stay in the emergency department for COVID-19 patients: A machine learning approach. *IEEE Access* 10:42243–42251.

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, ISSN 2291-9694.

Feder SL (2018) Data quality in electronic health records research: quality domains and assessment methods. *Western Journal of Nursing Research* 40(5):753–766.

Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A (2014) A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)* 46(4):1–37.

Gilboy N, Tanabe P, Travers DA, Rosenau AM, Eitel DR (2005) *Emergency Severity Index, Version 4: Implementation Handbook* (AHRQ Publication No. 05-0046-2. Rockville, MD).

Gill SD, Lane SE, Sheridan M, Ellis E, Smith D, Stella J (2018) Why do 'fast track'patients stay more than four hours in the emergency department? an investigation of factors that predict length of stay. *Emergency Medicine Australasia* 30(5):641–647.

Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. *International Conference on Machine Learning*, 1321–1330 (PMLR).

Hu Y, Cato KD, Chan CW, Dong J, Gavin N, Rossetti SC, Chang BP (2023) Use of real-time information to predict future arrivals in the emergency department. *Annals of emergency medicine* 81(6):728–737.

Hu Y, Chan CW, Dong J (2021) Prediction-driven surge planning with application in the emergency department. *Submitted to Management Science* .

Jones SS, Evans RS, Allen TL, Thomas A, Haug PJ, Welch SJ, Snow GL (2009) A multivariate time series approach to modeling and forecasting demand in the emergency department. *Journal of biomedical informatics* 42(1):123–139.

Karaca Z, Wong HS, Mutter RL (2012) Duration of patients' visits to the hospital emergency department. *BMC emergency medicine* 12:1–14.

Katayama Y, Kitamura T, Kiyohara K, Iwami T, Kawamura T, Izawa J, Gibo K, Komukai S, Hayashida S, Kiguchi T, et al. (2017) Improvements in patient acceptance by hospitals following the introduction of a smartphone app for the emergency medical service system: A population-based before-and-after observational study in osaka city, japan. *JMIR Mhealth and Uhealth* 5(9):e134–e134.

Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* 55(9):1486–1498.

Kharrazi H, Wang C, Scharfstein D (2014) Prospective EHR-based clinical trials: the challenge of missing data. *Journal of General Internal Medicine* 29:976–978.

King Z, Farrington J, Utley M, Kung E, Elkhodair S, Harris S, Sekula R, Gillham J, Li K, Crowe S (2022) Machine learning for real-time aggregated prediction of hospital admission for emergency patients. *NPJ Digital Medicine* 5(1):104.

Kwon H, Kim Y, Jo Y, Lee JH, Lee JB, Kim J, Hwang JE, Jeong J, Choi Y (2018) The korean triage and acuity scale: associations with admission, disposition, mortality and length of stay in the emergency department. *International Journal for Quality in Health Care* 31.

Lagoe RJ, Jastremski MS (1990) Relieving overcrowded emergency departments through ambulance diversion. *Hospital topics* 68(3):23–27.

Lee Y, Chen AS, Tajwar F, Kumar A, Yao H, Liang P, Finn C (2022) Surgical fine-tuning improves adaptation to distribution shifts.

Lin CH, Kao CY, Huang CY (2015) Managing emergency department overcrowding via ambulance diversion: A discrete event simulation model. *Journal of the Formosan Medical Association* 114(1):64–71.

McCoy TH, Pellegrini AM, Perlis RH (2018) Assessment of time-series machine learning methods for forecasting hospital discharge volume. *JAMA Network Open* 1(7):e184087–e184087.

Naeini MP, Cooper G, Hauskrecht M (2015) Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Nestor B, McDermott MBA, Boag W, Berner G, Naumann T, Hughes MC, Goldenberg A, Ghassemi M (2019) Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks. *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *PMLR*, 381–405 (PMLR).

Piccialli F, Giampaolo F, Prezioso E, Camacho D, Acampora G (2021) Artificial intelligence and healthcare: Forecasting of medical bookings through multi-source time-series fusion. *Information Fusion* 74:1–16.

Riguzzi C, Hern HG, Vahidnia F, Herring A, Alter H (2014) The july effect: is emergency department length of stay greater at the beginning of the hospital academic year? *Western Journal of Emergency Medicine* 15(1):88.

Riley P (2019) Three pitfalls to avoid in machine learning. *Nature* 572(7767):27–29.

Schweigler LM, Desmond JS, McCarthy ML, Bukowski KJ, Ionides EL, Younger JG (2009) Forecasting models of emergency department crowding. *Academic Emergency Medicine* 16(4):301–308.

Stein WE, Dattero R (1985) Sampling bias and the inspection paradox. *Mathematics Magazine* 58(2):96–99.

Stone K, Zwiggelaar R, Jones P, Mac Parthaláin N (2022) A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLoS Digital Health* 1:1–38.

Sun Y, Teow KL, Heng BH, Ooi CK, Tay SY (2012) Real-time prediction of waiting time in the emergency department, using quantile regression. *Annals of Emergency Medicine* 60(3):299–308.

Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, Godtliebsen F (2021) Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics* 13(6):e1549.

Thomas WJ, Guire KE, Horvat GG (1997) Is patient length of stay related to quality of care? *Journal of Healthcare Management* 42(4):489–507.

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* 26:1574–1583.

Trevino J, Malik S, Schmidt M (2022) Integrating google trends search engine query data into adult emergency department volume forecasting: Infodemiology study. *JMIR Infodemiology* 2:e32386.

Tuominen J, Koivistoinen T, Kanniainen J, Oksala N, Palomäki A, Roine A (2023) Early warning software for emergency department crowding. *Journal of Medical Systems* 47(1):66.

Tuominen J, Pulkkinen E, Peltonen J, Kanniainen J, Oksala N, Palomäki A, Roine A (2024) Forecasting emergency department occupancy with advanced machine learning models and multivariable input. *International Journal of Forecasting* 40(4):1410–1420.

Wang K, Hussain W, Birge JR, Schreiber MD, Adelman D (2022) A high-fidelity model to predict length of stay in the neonatal intensive care unit. *INFORMS Journal on Computing* 34(1):183–195.

Wargon M, Guidet B, Hoang TD, Hejblum G (2009) A systematic review of models for forecasting the number of emergency department visits. *Emergency Medicine Journal* 26:395 – 399.

Whitt W, Zhang X (2019) Forecasting arrivals and occupancy levels in an emergency department. *Operations Research for Health Care* 21:1–18.

Xiao C, Choi E, Sun J (2018) Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* 25(10):1419–1428.

Yang G, Hu EJ (2020) Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522* .

Yoo HH, Ro YS, Ko E, Lee JH, Han Sh, Kim T, Shin TG, Kim S, Chang H (2023) Epidemiologic trends of patients who visited nationwide emergency departments: A report from the national emergency department information system (nedis) of korea, 2018–2022. *Clinical and Experimental Emergency Medicine* 10(Suppl):S1.

Zaheer M, Kottur S, Ravanbakhsh S, Poczos B, Salakhutdinov RR, Smola AJ (2017) Deep sets. *Advances in Neural Information Processing Systems* 30.

Zeltyn S, Marmor YN, Mandelbaum A, Carmeli B, Greenshpan O, Mesika Y, Wasserkrug S, Vortman P, Shtub A, Lauterman T, et al. (2011) Simulation-based models of emergency departments: Operational, tactical, and strategic staffing. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 21(4):1–25.

Zhou L, Zhao P, Wu D, Cheng C, Huang H (2018) Time series model for forecasting the number of new admission inpatients. *BMC Medical Informatics and Decision Making* 18:1–11.

Appendix A: Bias when Reusing Length of Stay Predictions for Remaining Length of Stay Estimation

We use the notations from Section 2.3. For each patient i, their length of stay (LoS) is denoted λ_i and defined as $\lambda_i := \sum_t \delta_{i,t}$. For every time t such that the patient is in the ED at time t ($\delta_{i,t} = 1$) the remaining LoS for this patient is $r_i := \lambda_i - \sum_{s \leq t} \delta_{i,s}$. In this section, we show that, under mild conditions, a patient-level model designed to predict LoS leads to systematically biased estimates of the remaining LoS. Formally, we prove the following:

Lemma 2. *Consider a patient-level model for length of stay $\hat{\lambda}_i := \varphi(\boldsymbol{x}_i^{patient})$ satisfying*

$$\sum_i \hat{\lambda}_i \left(\lambda_i - \hat{\lambda}_i\right) = 0 \qquad \textit{(orthogonal residuals).}$$

Then, predicting $\hat{r}_i = \hat{\lambda}_i - \sum_{s \leq t} \delta_{i,s}$ for the remaining lenght of stay of patient i at time t is positively biased:

$$\sum_{(i,t):\delta_{i,t}=1} (r_{i,t} - \hat{r}_{i,t}) = \sum_i (\lambda_i - \hat{\lambda}_i)^2 \geq 0.$$

Proof The proof follows from simple algebraic manipulations. Indeed,

$$\sum_{(i,t):\delta_{i,t}=1} (r_{i,t} - \hat{r}_{i,t}) = \sum_{(i,t):\delta_{i,t}=1} \left(\lambda_i - \hat{\lambda}_i\right) = \sum_i \lambda_i \left(\lambda_i - \hat{\lambda}_i\right) = \sum_i \left(\lambda_i - \hat{\lambda}_i\right)^2 + \sum_i \hat{\lambda}_i \left(\lambda_i - \hat{\lambda}_i\right),$$

and the last term is zero, by assumption. □

The assumption in Lemma 2 requires the vector of predictions $(\hat{\lambda}_i)_{i=1,\ldots,N}$ to be orthogonal to the vector of residuals $(\lambda_i - \hat{\lambda}_i)_{i=1,\ldots,N}$. In particular, it follows naturally from the interpretation of ordinary least square as a projection and is satisfied by linear regression (see discussion in Appendix A.1). In addition, we see from Lemma 2 that the prediction bias is strictly positive as long as the length of stay model is not perfectly accurate, i.e., always in practice.

Intuitively, our analytical model highlights that bias can emerge due to the mismatch between the definition of an ‘observation’ between the two prediction tasks. For predicting LoS, each patient contributes equally, while, for predicting remaining LoS, each patient contributes proportionately to their length of stay—a discrepancy which is reminiscent of the inspection paradox in applied probability (Stein and Dattero 1985).

A.1. Orthogonal residual assumption

In this section, we elicit some generic conditions for the assumptions of Lemma 2, namely the orthogonal residuals assumption, to be satisfied.

We say that a family $\mathcal{F}$ of machine learning models is a *cone* if $\forall f \in \mathcal{F}$ and $\forall \alpha > 0$ we have that $\alpha f \in \mathcal{F}$. For example, the class of linear models (i.e., of the form $\boldsymbol{x} \mapsto \boldsymbol{\beta}^\top \boldsymbol{x}$ for some vector $\boldsymbol{\beta}$) or the class of linear models in a nonlinear feature space (i.e., of the form $\boldsymbol{x} \mapsto \boldsymbol{\beta}^\top \phi(\boldsymbol{x})$) are cones.

Lemma 3. *Consider a length of stay prediction model, $\hat{f}(\boldsymbol{x})$, obtained by solving an ordinary least square problem*

$$\min_{f \in \mathcal{F}} \quad \frac{1}{2} \sum_{i=1}^N (\lambda_i - f(\boldsymbol{x}_i))^2,$$

where the model class $\mathcal{F}$ is a cone. Then, the predictions $\hat{\lambda}_i := \hat{f}(\boldsymbol{x}_i)$ satisfy $\sum_{i=1}^N (\lambda_i - \hat{\lambda}_i)\hat{\lambda}_i = 0$.

Proof We denote $a := \sum_{i=1}^N (\lambda_i - \hat{\lambda}_i)\hat{\lambda}_i$ and show that $a = 0$ by contradiction. If $a \neq 0$, consider a new model $f_\eta = (1+\eta)f$ for some small $\eta \in (-1,1)$. We have $f_\eta \in \mathcal{F}$ since $\mathcal{F}$ is in a cone. Its error is given by

$$\sum_{i=1}^N (\lambda_i - f_\eta(\boldsymbol{x}_i))^2 = \sum_{i=1}^N \left(\lambda_i - \hat{\lambda}_i - \eta\hat{\lambda}_i\right)^2 \quad = \sum_{i=1}^N (\lambda_i - \hat{\lambda}_i)^2 - 2\eta \underbrace{\sum_{i=1}^N (\lambda_i - \hat{\lambda}_i)\hat{\lambda}_i}_{a} + \eta^2 \sum_{i=1}^N \hat{\lambda}_i^2.$$

Hence, if $a \neq 0$, taking $\eta \to 0$ with $a\eta > 0$ leads to a model f_η with $\sum_i (\lambda_i - f_\eta(\boldsymbol{x}_i))^2 < \sum_i (\lambda_i - \hat{\lambda}_i)^2$, which violates the optimality of $\hat{f}$. □

A.2. Unbiased predictions

In this section, we show that regression models trained with the sum of squared errors as the loss function achieve zero in-sample bias, for a wide class of predictive models.

Lemma 4. *Consider a length of stay prediction model, $\hat{f}(\boldsymbol{x})$, obtained by solving an ordinary least square problem*

$$\min_{f \in \mathcal{F}} \quad \frac{1}{2} \sum_{i=1}^N (\lambda_i - f(\boldsymbol{x}_i))^2 ,$$

where the model class $\mathcal{F}$ satisfies the following condition: $\forall f \in \mathcal{F}$, $\forall c \in \mathbb{R}$, $f + c \in \mathcal{F}$.

Then, the predictions $\hat{f}(\boldsymbol{x}_i)$ are unbiased, i.e., $\displaystyle\sum_{i=1}^N \left(\lambda_i - \hat{f}(\boldsymbol{x}_i)\right) = 0$.

Proof Denote b the value of the left-hand side. And consider $g := f + b/N \in \mathcal{F}$.

$$\sum_{i=1}^N (\lambda_i - g(\boldsymbol{x}_i))^2 = \sum_{i=1}^N \left(\lambda_i - \hat{f}(\boldsymbol{x}_i) - b/N\right)^2 = \sum_{i=1}^N \left(\lambda_i - \hat{f}(\boldsymbol{x}_i)\right)^2 - \left(\frac{b}{N}\right)^2.$$

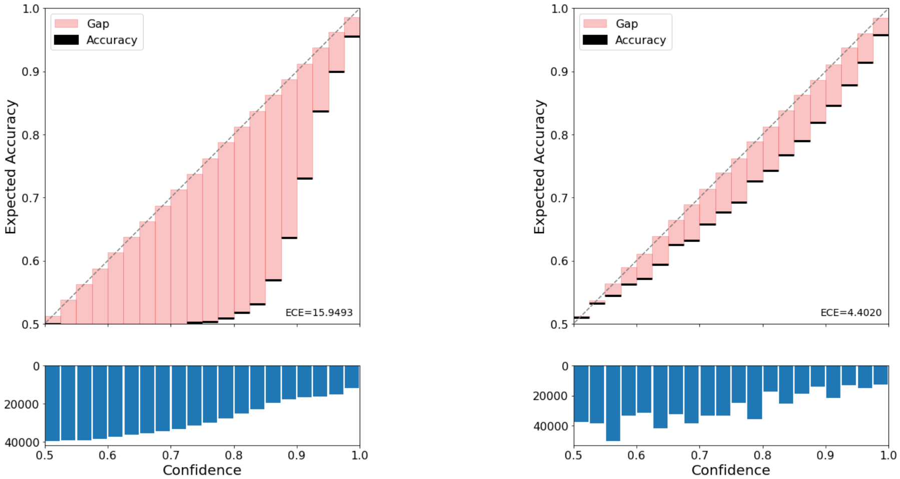
So, if $b \neq 0$, g achieves a strictly lower squared error than $\hat{f}$, hence contradicting its optimality. □

Appendix B: Additional Numerical Results

B.1. Aggregation of Patient-Level Predictions

In this section, we provide additional results on the accuracy of the patient-level predictive models developed in Section 4.1.

Figure 4 displays the calibration plots, on the training set, of the NN classifier, before and after the isotonic regression correction. Figure B.1 displays the same calibration diagrams (without any calibration and with the calibration fitted from the training set), on the testing set instead. In line with our discussion on nonstationarity and decalibration in Section 6, we observe that the calibration behavior of the raw scores differs on the test set compared to the training set. For example, we observe that the model with isotonic regression has (on the testing set) more mass at the higher levels of confidence and is generally over-confident on the test set compared to the base model. In particular, while effective on the training set, the isotonic regression fitted from the training data is not effective on the test set (and the resulting ECE score on the test set). This observation further motivates the need for a continuous (or rolling) calibration of the model across time. These observations are also valid for the XGB and RF models (calibration diagrams omitted).



(a) Without isotonic regression calibration (b) With isotonic regression calibration

Figure B.1 Calibration diagrams (on the testing set) of the NN classifier for predicting whether the remainig LoS of a given patient will exceeds four hours

Since Section 4 focuses on the bottom-up approach and the bias emerging (in-sample) from the misalignment between the patient-level and the system-level predictions, it reports accuracy metrics on the training set only and not out-of-sample performance (which is the focus of Section 6). For completeness, we report the accuracy of the three classification models (RF, XGB, NN), with and without isotonic regression, on the test sets, in Table B.1. Observe that the RF models had higher AUC scores than the XGB models on the training set (Table 3), but have slightly lower AUC on the test set. The ECE of the calibrated XGB model is lower (i.e., better) than the calibrated RF model on both the training and test sets.

Section 6 compares the out-of-sample performance of models from the time-series, bottom-up, and direct approaches respectively. In terms of bottom-up models, we only report the performance of classification models, with XGB and NN as classifiers. For the sake of completeness, we report the performance of one regression model (log(rLoS) predicted with RF) and the third classification model (RF) in Table B.2.

Table B.1 Accuracy metrics (on the test set) of different patient-level classification models (rLoS > 4 h) and their corresponding accuracy on occupancy prediction.

Task		RF		XGB		NN	
		raw	+ isotonic	raw	+ isotonic	raw	+ isotonic
Patient-level	AUC	0.734	0.734	0.747	0.747	0.724	0.724
	ECE	0.72	6.23	2.95	3.91	15.9	4.40
Occupancy	MAE	3.92	3.93	3.82	3.85	4.25	4.09
	MAPE	13.3%	12.5%	12.7%	12.7%	13.9 %	13.3%
	R^2	0.813	0.855	0.847	0.850	0.804	0.816
	Bias	1.55	0.485	2.05	2.02	2.19	1.26

Table B.2 Out-of-sample accuracy for predicting remaining occupancy with two bottom-up models: one regression model for log(rLoS)and one classification model. The regression model is calibrated using a simple linear regression step after aggregation and the classification model is calibrated with isotonic regression. We report performance without and with monthly recalibration. Models are trained on 2016-2018 data and evaluated on data from 2019.

Approach	Bottom-Up			
Patient-level outcome	log(rLoS)–regression		rLoS > 4h–classification	
Model	RF		RF	
Recalibration		✓		✓
MAE	5.37	4.46	3.93	3.73
MAPE	16.9%	14.7%	12.5%	12.0%
R^2	0.785	0.797	0.855	0.828
Bias	-1.02	-0.04	0.49	-0.09

B.2. Direct approach and deep set estimators

Table B.3 reports the in-sample performance of a deep set estimator with the architecture $\boldsymbol{\phi}: p$–64–d and $\rho: d$–d–1 for $d = 64$, as well as two comparable architectures that mimic bottom-up models (hence, with a latent dimension of 1).

Figure 5(a) displays the distribution of the number of nonzero coordinates in $\boldsymbol{\phi}(\boldsymbol{x})$. To complement these results, Figure B.2 reports, for each of the 64 coordinates of the embedding, how often it contains a nonzero value. We observe that four dimensions are very often nonzero, which correspond to the four nonzero coordinates of $\boldsymbol{\phi}(\mathbf{0})$. We also observe that some of the dimensions are essentially useless and never have a nonzero value. This suggests that a lower dimensional embedding could have been used and achieve similar performance.

Accordingly, we evaluate the performance of our architecture ($\boldsymbol{\phi}: p$–64–d and $\rho: d$–d–1) for various values of d in Figure B.3. We also experiment with the number of hidden nodes in the first layer of $\boldsymbol{\phi}$ (64, 32, 8). We observe that increasing the size of the latent space (hence, increasing the nonlinearity of the model) generally improves performance with a sharp gain as soon as $d \geq 8$. For consistency and simplicity, we keep $d = 64$ in our analysis but we note that the performance is fairly constant for values of d between 8 and 64.

Table B.3 In-sample results for predicting remaining occupancy using direct methods neural network with latent dimension of 1 and 64

Latent dimension	$d = 64$	$d = 1$	$d = 1$
$\boldsymbol{\phi}$	p–64–64	p–64–64–1	p–64–64–64–1
ρ	64–64–1	1–1	1–1
# Parameters	9,601	5,443	9,603
# Layers	4	4	5
MAE (patients)	2.78	2.79	**2.57**
MAPE	**62.4%**	63.2%	63.3%
R^2	0.949	0.949	**0.956**
Training Time (min)	6.24	5.13	6.14

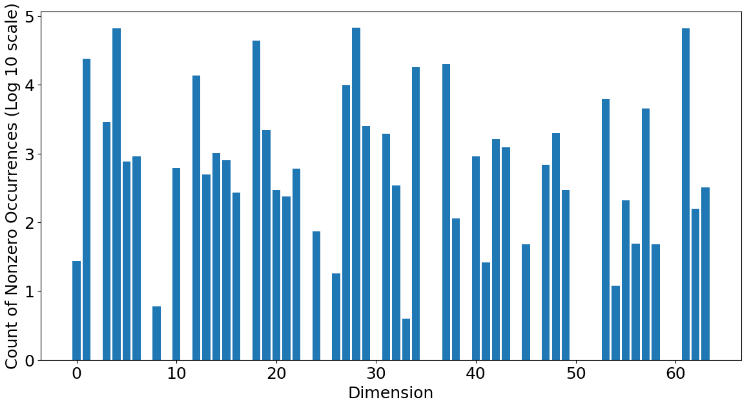
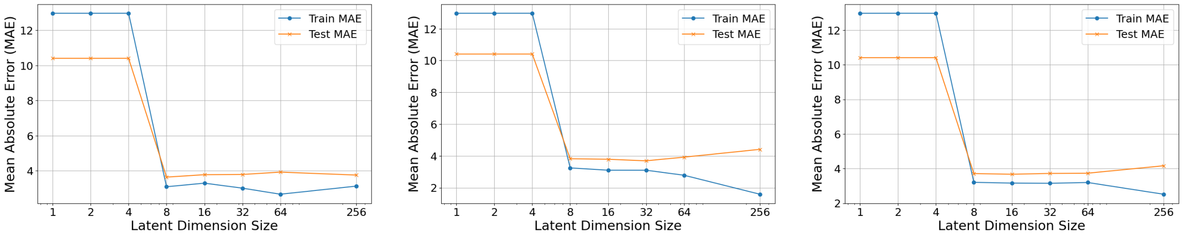


Figure B.2 **Count of the number of times each dimension of $\phi(x)$ is nonzero. Results are obtained by sampling 1,000 random time points t.**



(a) $k = 64$

(b) $k = 32$

(c) $k = 8$

Figure B.3 **In- and out-of-sample MAE for deep set estimators with architectures of the form $\phi : p–k–d$ and $\rho : d–d–1$**

We believe this behavior is due to the fact that ρ is also nonlinear, so nonlinearities in ρ can make up lower dimensional embedding.

To confirm this intuition, Figure B.4 represents the same metrics for varying d, yet with a highly nonlinear ρ, i.e., with a hidden layer of size 64 for all values of d ($\rho : d$–64–1). We observe a significantly different behavior with respect to d, with all of the previously observed variation absorbed by the additional complexity of ρ.

Figure B.5 represents the distribution of the relative predicted occupancy $\hat{z}_{t+1}/z_{t+1}$ as a function of the true remaining occupancy z_{t+1}. Overall, we observe that the distribution is concentrated around the value 1 (perfect prediction) but that it is more dispersed as the value of z_{t+1} decreases. In other words, the relative accuracy of the model is worse at low-occupancy values (≤ 20 beds) than at high ones. Although low-occupancy periods are not common (as indicated by the cooler values), the magnitude of the errors is such that it can have a noticeable effect on the overall MAPE. We should note, however, that the models tend to overestimate future occupancy ($\hat{z}_{t+1}/z_{t+1} > 1$), especially in this region. In practice, healthcare practitioners might be more comfortable with a model that overestimates occupancy (and incentivizes them to be more conservative) than underestimates.

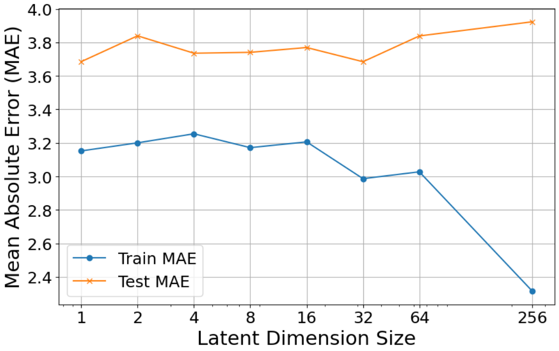


Figure B.4 **In- and out-of-sample MAE for deep set estimators with architectures of the form $\phi: p$–64–d and $\rho: d$–64–1**

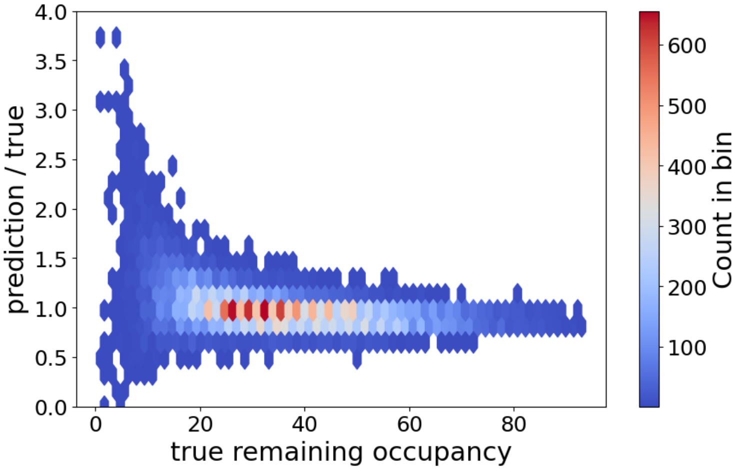


Figure B.5 **Scatter plot of the remaining occupancy z_{t+1} (x-axis) vs. the relative predicted occupancy $\hat{z}_{t+1}/z_{t+1}$ (y-axis) for the deep set estimator, on the training dataset.**

Occupancy Prediction with Patient Data: Evaluating Time-Series, Patient-Level Aggregation, and Deep Set Models
Electronic Companion

Appendix EC.1: Methods: Prediction Models and Metrics

EC.1.1. Random forests

Random forests are a popular and effective machine learning method for regression and classification tasks (Breiman 2001). They are an example of ensemble methods, which are models consisting of a group of weak learners that are combined to create a strong model.

In a random forest, each weak learner is a decision tree. Each tree in the random forest is trained on a different subset of the data, and the subsets are chosen randomly with replacement. This process is known as bootstrapping. Additionally, a random subset of the overall available features in the data is chosen for each tree to learn on. To make a prediction, the random forest averages the predictions of all the decision trees in the forest. For classification tasks this amounts to selecting the label with the highest number of votes by all the trees. Crucially we can also obtain a notion of predicted probability for each class from the random forest classifier by measuring the fraction of trees voting for a given class.

In terms of hyper-parameters, we consider the following in our implementation: For classification models, we tune the maximum depth of each tree, the maximum number of features considered by each tree, and the maximum number of samples used by each tree. For regression models we use the `scikit-learn` default parameters (Pedregosa et al. 2011) except for the maximum number of features considered (whose default parameter is $\sqrt{p}$ where p is the total number of available features), which we set to be a third of the number of features, $p/3$, as recommended in Hastie et al. (2009, Chapter 15). We calibrate hyper-parameters using grid search and k-fold ($k = 4$) cross-validation on the training set.

EC.1.2. Gradient boosting

Gradient boosting is a machine learning method for regression and classification problems, and is another example of an ensemble method. Similar to random forests, the weak learners in a gradient boosting model are decision trees. However, in gradient boosting, the model starts with a base learner and then in each successive step, a new weak learner is trained to correct the mistakes of the incumbent model. The final model is a weighted sum of the individual weak models, where the weights are learned using the gradient descent algorithm.

XGBoost (eXtreme Gradient Boosting) is a popular library for efficiently training powerful gradient boosting models (Chen and Guestrin 2016). XGBoost has been shown to outperform other standard ensemble methods on both clinical prediction tasks (Chang et al. 2019, Wang et al. 2020) and operational tasks such as predicting number of hospital admissions (King et al. 2022).

The hyper-parameters of the XGB models are the number of trees in the ensemble, the learning rate, and the maximum depths of the trees. We calibrate hyper-parameters using grid search and k-fold ($k = 4$) cross-validation on the training set.

EC.1.3. Neural Networks

Neural networks are a class of nonlinear models that are the cornerstone of deep learning (Goodfellow et al. 2016).

A neural network can be described as a series of layers, with each layer consists of units (or neurons). Denoting $\boldsymbol{z}_\ell$ the output vector of layer ℓ, each unit of layer $\ell + 1$ returns an output of the form $f(\boldsymbol{w}^\top \boldsymbol{z}_\ell + b)$, where f is called an activation function (we take the ReLu function $f(t) = \max(0, t)$ in our experiments) and where $\boldsymbol{w}/b$ are unit-specific weights/intercept. For simplicity, we consider architectures where layer ℓ serves as an input for (or feeds) layer $\ell + 1$ only and where all units from layer ℓ are connected to all units in layer $\ell + 1$ (fully connected). Hence, the architecture of the network can be summarized by the size (i.e., the number of units) in each layer, starting from the input layer (corresponding to the input vector $\boldsymbol{x} \in \mathbb{R}^p$). For example, linear/logistic regression models can be represented by neural networks of the form p–1; a neural network with 2 hidden layers of size 25 and predicting a scalar output is of the form p–25–25–1.

One of the main drivers for the broad adoption of neural networks in practice is the development of open-source frameworks to define and train neural networks efficiently, such as TensorFlow or PyTorch. We use TensorFlow in our experiments.

In terms of architecture, we do our best to compare architectures with the same number of parameters across the three approaches (namely, $9,600$). For the time-series and bottom-up approach, we consider architectures of the form p–64–64–64–1. For the direct approach, our focal architecture is $\boldsymbol{\phi} : p$–64–64 and $\rho : 64$–64–1.

Every model is trained using the Adam optimizer, a batch size of 64, and a learning rate of 0.001. To avoid overfitting, models are trained for up to 1,000 epochs with the mean squared error (or the binary cross entropy loss, for classification) on a holdout validation set used as an early stopping criterion (patience parameter of 25 epochs).

EC.1.4. Isotonic regression

A flexible, nonparametric method to calibrate the output probabilities of a binary classifier is to partition the predicted probabilities into bins, p_b, and to estimate an empirical probability q_b for each such bin. We then replace the predicted probability p_b with q_b. This general approach is known as histogram binning (Zadrozny and Elkan 2001, Murphy 2023).

To avoid overfitting, we can replace the predicted probabilities p_b by $f(p_b)$ for some well-structured function f instead. For example, we can calibrate f so that it is monotically nondecreasing and fits q_b as closely as possible, i.e., solve

$$\min_{\hat{q}_b} \sum_b (q_b - \hat{q}_b)^2 \quad \text{s.t.} \quad \hat{q}_b \leq \hat{q}_c \text{ if } p_b \leq p_c \text{ for all pairs of bins } (b,c).$$

This calibration approach is specifically referred to as isotonic regression (Zadrozny and Elkan 2001, Murphy 2023) and is the one we use in this paper.

Alternatively, we can find f by fitting a logistic model to predict q_b as a function p_b.

EC.1.5. Performance metrics

Our final prediction task (predicting remaining occupancy in the ED), can be seen as a regression task. Accordingly, we measure prediction error using different metrics commonly used in regression: the mean absolute error (MAE) and the mean absolute percentage error (MAPE). Formally, if y_i denote the target outcome for sample $i = 1, \ldots, n$, and $\hat{y}_i$ denote our prediction, MAE and MAPE are defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad \text{and} \quad MAPE = \frac{100}{n}\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|.$$

We also report the proportion of variance explained by the model, i.e., R^2, defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{with} \quad \bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i.$$

When comparing patient-level binary classification models, we measure accuracy using the area under the receiver operating characteristic curve (AUC), which also represents the probability of correctly prioritizing two random samples with label 0 and 1 respectively. The AUC can take any value from 0.5 to 1, with 1 indicating a perfect classifier.

For classification models, another important property is calibration. A model is said to be 'calibrated' when the predicted probabilities match the empirical frequency of event occurrence (Guo et al. 2017). A common calibration measure is the Expected Calibration Error (ECE; see Guo et al. 2017, Naeini et al. 2015). Consider the same notations as above and further assume that $y_i \in \{0,1\}$ and $\hat{y}_i \in [0,1]$. To compute the ECE, we discretize the interval $[0,1]$ into B bins of the form $(\frac{b-1}{B}, \frac{b}{B}]$. Let $\mathcal{I}_b \subseteq \{1, \ldots, n\}$ denote the set of observations in bin b, i.e., the samples i such that the predicted probability $\hat{y}_i$ belongs to $(\frac{b-1}{B}, \frac{b}{B}]$. Then, for each bin, we define the accuracy and confidence of the bin as

$$\mathrm{acc}(\mathcal{I}_b) := \frac{1}{|\mathcal{I}_b|}\sum_{i \in \mathcal{I}_b} y_i, \quad \text{and} \quad \mathrm{conf}(\mathcal{I}_b) = \frac{1}{|\mathcal{I}_b|}\sum_{i \in \mathcal{I}_b} \hat{y}_i.$$

In other words, the accuracy and confidence correspond to the empirical average of y_i and $\hat{y}_i$ respectively. The ECE is defined as the weighted average of the absolute difference between accuracy and confidence:

$$ECE := \sum_{b=1}^{B} \frac{|\mathcal{I}_b|}{n} \left|\mathrm{acc}(\mathcal{I}_b) - \mathrm{conf}(\mathcal{I}_b)\right|.$$

Visually, the accuracy $\mathrm{acc}(\mathcal{I}_b)$ can be represented as a function of the confidence $\mathrm{conf}(\mathcal{I}_b)$—see Figure 4 for example—and the ECE measures the distance between this curve and the diagonal, which corresponds to perfect calibration. We refer to Murphy (2023) for a generalization of ECE for multi-class classification.

Appendix EC.2: Analytical Comparison of Distribution Shift Detection at Patient and Observation Levels

Distribution shift, which generally refers to a change in the data distribution between the training and testing phases, is a serious hurdle in the adoption of machine learning systems in healthcare (Challen et al. 2019). The first step involves the efficient detection of distribution shifts (Vovk et al. 2021).

In this section, we illustrate the impact of the misalignment between the patient- and system-level point of view, on the definition and detection of nonstationarity. Indeed, most studies related to predictions of patient-level outcomes report summary statistics on=f patient-level covariates on the dataset where each patient contributes equally. However, as we illustrated in a previous analysis (Section 2.4), for system-level prediction, each patient contributes proportionately to their length of stay. Here, we show analytically that a change can be considered statistically significant at a patient level and become insignificant for system-level prediction, or vice versa.

We assume that we have access to a single patient covariate $x_i^{\mathtt{patient}} \in \mathbb{R}$ (which we later denote x_i for concision) which can be used to predict length of stay. We denote the length of stay of patient i, λ_i, which we consider here as a continuous variable. For simplicity, we assume that (x_i, λ_i) follows a multivariate normal distribution, whose parameters are unknown. We denote μ_x, σ_x^2 (resp $\mu_\lambda, \sigma_\lambda^2$) the mean and variance of x (resp. λ) and let $\rho \in [-1, 1]$ denote the correlation factor between x_i and λ_i. Note that we assume $\mu_x, \sigma_x^2, \mu_\lambda, \sigma_\lambda^2, \rho$ are all unknown. Formally, we write

$$\begin{pmatrix} x_i \\ \lambda_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_\lambda \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_\lambda \\ \rho\sigma_x\sigma_\lambda & \sigma_\lambda^2 \end{pmatrix} \right).$$

We consider a simple example of distributional shift: We assume that we have two time periods, which we refer to as the training and test set respectively, with the training period ending at time $T^{\mathtt{train}}$. Assume that the mean of x_i shifts from μ_x to $\mu_x + \varepsilon$ after $T^{\mathtt{train}}$ (everything else remaining equal). Let $S_1 = \{i : \exists t \leq T^{\mathtt{train}} \text{ for which } \delta_{it} = 1\}$ be the patients in the training set and $S_2 = \{1,\dots,N\} \setminus S_1$ be the patients in the test set. Define $N_1 := |S_1|$ and $N_2 := |S_2|$. Let $\overline{X}_1 = \frac{1}{N_1}\sum_{i\in S_1} x_i$ and $\overline{X}_2 = \frac{1}{N_2}\sum_{j\in S_2} x_j$ be our mean estimates in the respective time periods.

We investigate the ability of a t-test to detect the distribution shift. At a *patient* level, the t-test statistic is given by

$$T_p = \frac{\overline{X}_2 - \overline{X}_1}{S_p\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}}, \qquad \text{where } S_p^2 := \frac{1}{N_1+N_2-2}\Big(\sum_{i\in S_1}(x_i-\overline{X}_1)^2 + \sum_{j\in S_2}(x_j-\overline{X}_2)^2\Big)$$

is the pooled sample variance.

However, for system-level predictions, the relevant dataset is one where each patient is weighted proportionately to their length of stay. This is equivalent to detecting a shift in the average of $\lambda_i x_i$ instead. Although only the mean of x_i changes, note that both the mean and variance of $\lambda_i x_i$ vary. Let $\overline{Y}_1 = \frac{1}{N_1}\sum_{i\in S_1}\lambda_i x_i$ and $\overline{Y}_2 = \frac{1}{N_2}\sum_{j\in S_2}\lambda_j x_j$. Accordingly, the associated t-statistic, at the *observation* level, is given by

$$T_o = \frac{\overline{Y}_2 - \overline{Y}_1}{S_o\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}} \qquad \text{where } S_o^2 := \frac{1}{N_1+N_2-2}\Big(\sum_{i\in S_1}(\lambda_i x_i-\overline{Y}_1)^2 + \sum_{j\in S_2}(\lambda_j x_j-\overline{Y}_2)^2\Big).$$

Then, as $N_1, N_2 \to +\infty$ with $N_1/N_2 \to c$ for some $0<c<\infty$, we have $T_p \to_d t_p$ and $T_o \to_d t_o$ with

$$t_p := \frac{\varepsilon}{\sigma_x\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}},$$

$$t_o := \frac{\varepsilon\mu_\lambda}{s\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}} \text{ with } s := \sqrt{\frac{(N_1-1)\mathrm{Var}(\lambda_i x_i) + (N_2-1)\mathrm{Var}(\lambda_j x_j)}{N_1+N_2-2}},$$

where x_i, λ_i (resp. x_j, λ_j) are distributed according to the distribution of (x,λ) on the training set (resp. test set).

If $t_p > t_o$, then a change of ε can be considered as significant for the patient-level test, while remaining undetected at a patient-time level. The reverse situation occurs when $t_p < t_o$. As clear from the analytical expression above, comparing t_o with t_p boils down to comparing s with $\mu_\lambda\sigma_x$. Depending on the coefficient of variations of x and λ and that of their correlation coefficient ρ, this difference can be negative or positive, as we formally state in the special case where $\varepsilon \ll \mu_x$.

Lemma EC.1. *In the regime where $\varepsilon \ll \mu_x$, we have*

$$s^2 - (\mu_\lambda\sigma_x)^2 \approx \sigma_x^2\sigma_\lambda^2\left\{\rho^2 + 2\rho\frac{\mu_x}{\sigma_x}\frac{\mu_\lambda}{\sigma_\lambda} + (\frac{\mu_x^2}{\sigma_x^2}+1)\right\}.$$

In addition, the second-order polynomial in ρ on the right-hand side

- *is positive for $\rho = 0$,*
- *takes negative values on the interval $[-1,1]$ if and only if $2 - 2\left|\frac{\mu_x}{\sigma_x}\frac{\mu_\lambda}{\sigma_\lambda}\right| + \frac{\mu_x^2}{\sigma_x^2} < 0$.*

From the first condition in Lemma EC.1, we conclude that $t_o < t_p$ whenever $\rho = 0$, meaning that a change ε in the average value of covariate x can be statistically significant at a patient level but insignificant at a (patient, time) one. Conversely, one can easily find parameter values satisfying the second condition (e.g., by fixing μ_x, σ_x, and σ_λ and taking μ_λ large enough), where it is possible to have $t_o > t_p$, i.e., detectable nonstationarity at an observation level but not at a patient level. Actually, this situation occurs in our dataset for 2 out of 8 patient covariates (see Online Supplement EC.3.2).

Proof of Lemma EC.1 Following Craig (1936) and Haldane (1942) we have that the first two moments of $x_1\lambda_1$ and $x_2\lambda_2$ are given by

$$\begin{aligned}
\mathbb{E}(x_1\lambda_1) &= \mu_x\mu_x + \rho\sigma_x\sigma_\lambda,\\
\mathbb{E}(x_2\lambda_2) &= (\mu_x+\varepsilon)\mu_\lambda + \rho\sigma_x\sigma_\lambda = \mathbb{E}(x_1\lambda_1) + \varepsilon\mu_\lambda,\\
\mathrm{Var}(x_1\lambda_1) &= \mu_x^2\sigma_\lambda^2 + \mu_\lambda^2\sigma_x^2 + 2\rho\mu_x\mu_\lambda\sigma_x\sigma_\lambda + (1+\rho^2)\sigma_x^2\sigma_\lambda^2,\\
\mathrm{Var}(x_2\lambda_2) &= (\mu_x+\varepsilon)^2\sigma_\lambda^2 + \mu_\lambda^2\sigma_x^2 + 2\rho(\mu_x+\varepsilon)\mu_\lambda\sigma_x\sigma_\lambda + (1+\rho^2)\sigma_x^2\sigma_\lambda^2\\
&= 2\varepsilon\mu_x\sigma_\lambda^2 + \varepsilon^2\sigma_\lambda^2 + 2\rho\varepsilon\mu_\lambda\sigma_x\sigma_\lambda + \mathrm{Var}(x_1\lambda_1),\\
s^2 &= \mathrm{Var}(x_1\lambda_1) + \frac{N_2-1}{N_1+N_2-2}(2\varepsilon\mu_x\sigma_\lambda^2 + \varepsilon^2\sigma_\lambda^2 + 2\rho\varepsilon\mu_\lambda\sigma_x\sigma_\lambda).
\end{aligned}$$

If $\delta \ll \mu_x$, then $s^2 \approx \mathrm{Var}(x_1\lambda_1)$ and

$$\begin{aligned}
s^2 - (\mu_\lambda\sigma_x)^2 &\approx \mathrm{Var}(x_1\lambda_1) - (\mu_\lambda\sigma_x)^2\\
&= \mu_x^2\sigma_\lambda^2 + 2\rho\mu_x\mu_\lambda\sigma_x\sigma_\lambda + (1+\rho^2)\sigma_x^2\sigma_\lambda^2\\
&= \sigma_x^2\sigma_\lambda^2\left\{\rho^2 + 2\rho\frac{\mu_x}{\sigma_x}\frac{\mu_\lambda}{\sigma_\lambda} + (\frac{\mu_x^2}{\sigma_x^2}+1)\right\}\\
&=: \sigma_x^2\sigma_\lambda^2 P(\rho).
\end{aligned}$$

Obviously, $P(0) = \dfrac{\mu_x^2}{\sigma_x^2} + 1 > 0$.

Hence, $P(\rho)$ can take negative values if and only if it admit two real roots, i.e., if and only if its discriminant

$$\Delta := \left(\frac{\mu_x}{\sigma_x}\frac{\mu_\lambda}{\sigma_\lambda}\right)^2 - \left(\frac{\mu_x^2}{\sigma_x^2}+1\right)$$

is positive. In this case, $P(\rho)$ is negative over the open interval (ρ_-, ρ_+) with

$$\rho_\pm := -\frac{\mu_x}{\sigma_x}\frac{\mu_\lambda}{\sigma_\lambda} \pm \sqrt{\Delta}.$$

However, $\rho_-\rho_+ = \dfrac{\mu_x^2}{\sigma_x^2} + \geq 1 > 0$, which implies that ρ_- and ρ_+ have the same sign and that at least one of them as magnitude more than 1. Hence, (ρ_-, ρ_+) intersects $[-1, 1]$ if and only if $\min\{|\rho_-|, |\rho_+|\} < 1$.

Observe that $\sqrt{\Delta} < \left|\frac{\mu_x}{\sigma_x}\frac{\mu_\lambda}{\sigma_\lambda}\right|$ so the above condition is equivalent to

$$\begin{cases} \rho_- < 1 & \text{if } -\frac{\mu_x}{\sigma_x}\frac{\mu_\lambda}{\sigma_\lambda} \geq 0 \\ \rho_+ > -1 & \text{if } -\frac{\mu_x}{\sigma_x}\frac{\mu_\lambda}{\sigma_\lambda} < 0 \end{cases} \iff \sqrt{\Delta} > \left|\frac{\mu_x}{\sigma_x}\frac{\mu_\lambda}{\sigma_\lambda}\right| - 1.$$

As a result, we can say that $P(\rho)$ takes negative values on the interval $[-1,1]$ if and only if

$$\begin{cases} \Delta > 0 \\ \sqrt{\Delta} > \left|\frac{\mu_x}{\sigma_x}\frac{\mu_\lambda}{\sigma_\lambda}\right| - 1 \end{cases} \iff \Delta > \left(\left|\frac{\mu_x}{\sigma_x}\frac{\mu_\lambda}{\sigma_\lambda}\right| - 1\right)^2$$

$$\iff 2 - 2\left|\frac{\mu_x}{\sigma_x}\frac{\mu_\lambda}{\sigma_\lambda}\right| + \frac{\mu_x^2}{\sigma_x^2} < 0.$$

□

Appendix EC.3: Additional Data Statistics and Figures

In this section, we provide additional description of our data for ED occupancy prediction.

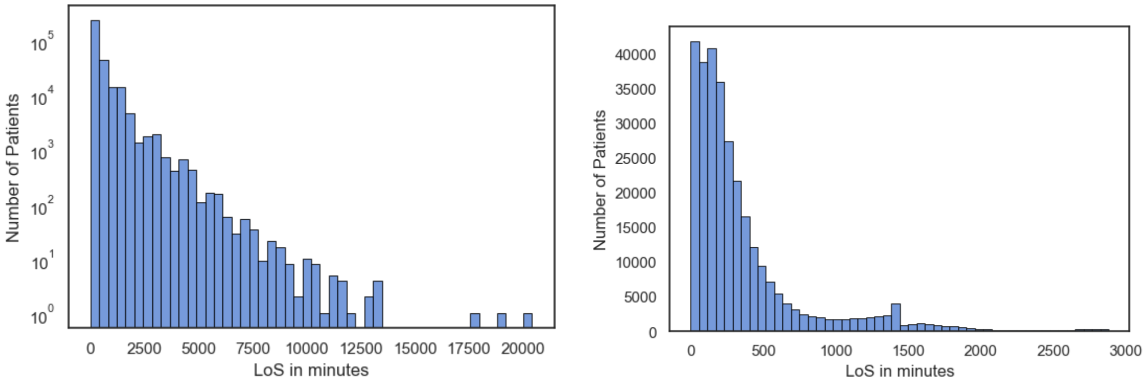
EC.3.1. Data description

We present summary statistics for the length of stay (LoS) in the ED and all patient-level features in Table EC.1. We observe that LoS varies greatly among patients in our dataset with a standard deviation of 643 minutes for a 423-minute mean. Figure EC.1 provides a histogram of the LoS of all patients visiting the ED during our study period (left panel, log-scale for the vertical axis) as well as a histogram of the LoS of all patients staying less than 48 hours. Note a bump in Figure EC.1(b) at 1,440 minutes, i.e., 24 hours. After discussion with our partner hospital, we believe this bump results from a directive at our partner hospital of keeping no more than 5% of patients in the ED for more than 24 hours.

The demographic feature columns (Female and Age) show that about 51% of the patients were female and the mean age of all patients was about 45 years old, with a a standard deviation of 25 years. The range of recorded values for vital sign measurements were validated as reasonable by our partner hospital. The median KTAS level was 4. Note that primary symptom code is excluded from Table EC.1 due to being an unordered categorical variable.

Table EC.1 Summary statistics and percentage missing for length of stay and patient-level features

	LoS (min)	Female	Age	SBP	DBP	PR	RR	Temp (°C)	SpO2	KTAS
Mean	422.62	0.51	45.46	110.62	65.86	90.50	18.95	37.04	97.55	3.50
STD	642.96		25.35	50.38	30.23	21.77	3.71	0.85	3.63	0.76
Q1	111.00		26.23	103.00	60.00	75.00	18.00	36.50	97.00	3.00
Median	224.00		50.08	121.00	73.00	87.00	18.00	36.90	98.00	4.00
Q3	434.00		65.42	140.00	84.00	102.00	20.00	37.40	99.00	4.00
% Missing	0.0	0.0	0.0	2.4	2.4	14.0	14.0	17.0	2.3	3.0



(a) All patients

(b) Patients staying less than 48 hours

Figure EC.1 Distribution of length of stay in the ED between January 9, 2016 and December 31, 2019

Table EC.2 Distribution of KTAS accross the study population and summary statistics on LoS for each KTAS category

KTAS		LoS		
Value	Frequency	Median	Mean	Std. Dev.
Missing	0.03	55.00	91.46	100.06
Level 1	0.01	129.00	360.40	610.88
Level 2	0.06	317.00	542.03	705.42
Level 3	0.41	306.00	543.95	722.38
Level 4	0.42	182.00	342.13	564.01
Level 5	0.07	110.00	240.82	488.75

Table EC.1 also reports the fraction of missing entries for each patient-level feature. Demographic information (Female and Age) are systematically observed. Only 3% of patients or less are missing SBP, DBP, SpO2, KTAS, or primary symptom code. The features with the highest missing rates (14–17%) are PR, RR, and Temperature. Our strategy for handling missing values depends on the model being used. If the model can support input with missing entries directly (e.g., XGBoost), we keep the missing entries as missing. Otherwise (e.g., when using random forests), we replace missing numerical variables (e.g., vital signs) by the median and encode missing categorical variables (e.g., KTAS) by adding 'missing' as a new possible category as recommended in Bertsimas et al. (2024).

While Table EC.1 presents summary statistics on the LoS distribution in our full study population, Table EC.2 reports the median, average, and standard deviation of LoS stratified by KTAS value. In our data, the percentage of patients with LoS over four hours is 47%.

EC.3.2. Detecting distribution shifts

Figure 1 clearly indicates a shift in the distribution of our outcome of interest (here the probability of a patient currently in the ED stays at least four hours). This distribution shift could be due to nonstationarity in the patient population (covariate shifts) or in the relationship between the patient covariates and their LoS.

To assess the relevance of nonstationarity for our predictive task, we compare the distribution of eight of our patient-level variables on the training and the testing set. Formally, we consider all the variables presented in Table EC.1 and test whether their mean (t-test) or their distribution (Kolmogorov-Smirnov test) is the same between the training and the test set. We report the p-values of these tests in Table 3(a). In summary, for seven out of the nine variables, both tests conclude that the covariate distribution is different on the training and testing set, with a p-value lower than $5 \cdot 10^{-3}$. Hence, we can conclude that nonstationarity is a prominent issue in our data, even for demographic or clinical variables that could be thought as more stationary than operational characteristics of the ED. Age and sex are the only two variables for which we could not reject the null hypothesis for at least one of the tests.

To detect nonstationarity, we compare the covariates of patients in the training and test set. However, following the analysis from Section 2.4, we should keep in mind that each predictive task defines a set of 'observations' and that the definition of nonstationarity depends on this task. In particular, in our case study, we are interested in hourly predictions of patient outcomes, i.e., our models apply to a dataset where each patient contributes proportionately to their LoS. Accordingly, we should also test for nonstationarity in that dataset, as reported in Table 3(b). For Temperature and Female, we observe that, when weighting patients by their LoS (which, again, is the relevant weighting for our patient-level models), we detect statistically significant changes in the distributions of both variables, although their means on the patient data does not significantly change.

Table EC.3 Results from t- and Kolomogorov-Smirnov (KS) tests to detect nonstationarity in the distribution of our patient-level variables between the training and testing data.

(a) When patients are weighted equally

Variable	Average		t-test		KS-test	
	Training data	Test data	Statistic	p-value	Statistic	p-value
Age	45.05	46.52	-14.32	*	0.03	*
Temperature	37.04	37.04	-0.00	1.00	0.02	*
SBP	109.92	112.40	-12.11	*	0.03	*
DBP	65.37	67.11	-14.16	*	0.04	*
PR	89.84	92.09	-24.13	*	0.03	*
RR	18.97	18.91	3.73	0.0005	0.11	*
SpO2	97.49	97.69	-12.83	*	0.06	*
Female	0.51	0.51	0.37	0.71	0.00	1.00

Note: $*$: $< 10^{-4}$

(b) When each patient are weighted proportionately to their LoS

Variable	Average		t-test		KS-test	
	Training data	Test data	Statistic	p-value	Statistic	p-value
Age	54.17	53.79	11.16	*	0.02	*
Temperature	37.18	37.13	35.74	*	0.03	*
SBP	118.26	118.54	-5.13	*	0.01	*
DBP	69.95	70.17	-7.10	*	0.01	*
PR	94.29	94.13	7.08	*	0.02	*
RR	19.14	18.89	47.79	*	0.12	*
SpO2	97.10	97.34	-50.15	*	0.07	*
Female	0.47	0.49	-27.02	*	0.02	*

Note: $*$: $< 10^{-4}$

In addition to demographic and clinical features, our patient-level predictions also rely on the time already spent in the ED by each patient. By visual inspection, we observe in Figure EC.2 that this covariate also significantly varies throughout our study period. In particular the average time already spent for patients in the training set is 776 minutes, while it is 486 minutes only for patients in the test set.

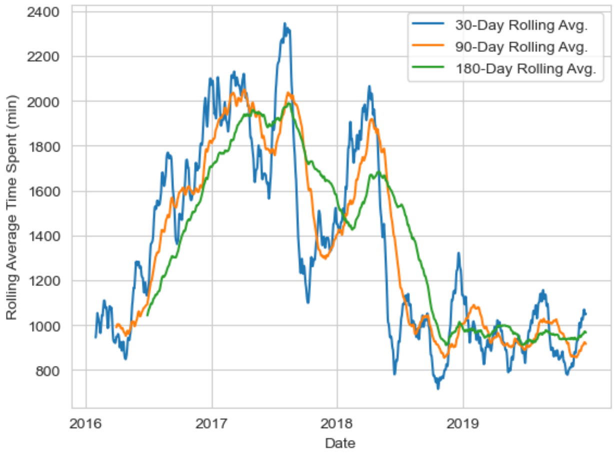


Figure EC.2 Evolution of the average (rolling average, for three different time windows) of the variable 'time spent in the ED' over our study period

Appendix EC.4: Literature on Patient-Level Length of Stay Estimation

There is an extensive literature on predicting LoS in emergency departments and other hospital units. We refer the reader to Awad et al. (2017), Stone et al. (2022), Bacchi et al. (2022) for recent surveys. Recent years have seen a growing interest for more sophisticated and sometimes black-box models, such as neural nets or ensemble methods, whose success compared with simpler interpretable models is viewed as "irrefutable" (Stone et al. 2022). In this section, we broadly review the use of three types of models in the LoS prediction literature: regression, classification, and survival models.

In empirical studies, LoS is used as a proxy for quality of care (see Thomas et al. 1997, Brasel et al. 2007, for discussions on its relevance) and is involved in linear regression analyses. Yet, the objective of these works is not to provide accurate estimates of LoS but rather identify causal factors and quantify the magnitudes of their impact on LoS. Since the distribution of LoS is usually right-skewed, some regression models apply to the logarithm of the LoS instead.

Discretizing LoS into categories is another alternative to address the right-skewedness of LoS distribution that has gained increased attention. Out of the 21 publications on predicting medical inpatient LoS reviewed by Bacchi et al. (2022), 12 of them consider classification models, with half of them being published after 2018, showing a growing interest for classification models for LoS prediction. We believe this trend parallels a growing interest for purely predictive models that achieve high accuracy.

Only one of the 21 reviewed papers simultaneously considers regression and classification models (Baek et al. 2018). Baek et al. (2018) build LoS predictions across all inpatient admissions of a tertiary general university hospital in South Korea. They achieve an MAE of 4.68 days and R^2 of

0.267 for the regression task using a linear regression model. They also construct a random forest classifier that predicts whether a patient stays more than 30 days with an AUC of 0.97. However, as Bacchi et al. (2022) point out, directly comparing regression and classification approaches solely on the basis of accuracy is difficult due to the absence of a common metric.

Besides regression and classification, length of stay prediction can also be studied with tools from survival analysis. In this context, a patient is "surviving" at a given time point if they are still present in the hospital or unit. We refer the reader to another survey on LoS modeling by Awad et al. (2017) that discusses survival models for the general hospital LoS prediction task. Chaou et al. (2017) consider the use of of survival models for ED patients in particular, but with a focus on identifying and quantifying the relative importance of factors determining LoS rather than attempting to optimize prediction accuracy. Similar studies such as Agarwal et al. (2021) and Wang et al. (2022b) identify characteristics of COVID-19 patients that are correlated with longer hospital lengths of stay. Wang et al. (2022a) adapt a survival forest approach for predicting LoS in neonatal ICU. We were unable to perform an exhaustive cross-validation of hyper-parameters for these models due to increased computational cost (especially for nonlinear models such as survival trees), which prevented us from making a fair comparison with regression and classification models. Accordingly, we decided not to include survival models in our analysis. Quantile regression models have also been used to provide probabilistic forecasts of patient LoS or waiting times (see, e.g., Arora et al. 2023).

We conclude this section by commenting on the recent literature focusing on ED patients specifically. Kadri et al. (2022) view predicting LoS in a pediatric ED as a regression problem; they use various nonlinear machine learning methods and obtain the highest accuracy ($R^2 = 0.871$) with generative adversarial networks. Benbelkacem et al. (2019) formulate the problem of predicting LoS in a pediatric ED as a classification task and compare the accuracy of several traditional supervised machine learning models, including naive Bayes, support vector machine, and decision trees. Close to our work, Gill et al. (2018) and Etu et al. (2022) train models to predict whether the LoS of ED patients exceeds 4 hours (note that the second study focuses on Covid-19 patients only) and achieve the best accuracy—AUC about 90%—with gradient boosted trees.

Appendix EC.5: Supplementary Data to Section 7

This section provides complementary information to the extensions presented in Section 7

EC.5.1. Performance across multiple prediction horizons

In Section 7.1, we evaluate the robustness of our findings to the prediction horizon. We replicate our main analysis (4-hour ahead prediction) to prediction horizons ranging from 1 to 12 hours. We compare out-of-sample MAE, MAPE, R^2, and bias in Table EC.1 and Figure EC.1.

Table EC.1 Comparison of MAE, MAPE, R^2, and bias for our three prediction approaches and for varying forecasting horizons (rounded to two decimals).

Hour	Time Series				Bottom-Up				Direct			
	MAE	MAPE	R^2	Bias	MAE	MAPE	R^2	Bias	MAE	MAPE	R^2	Bias
1	2.42	4.60	0.96	-0.15	2.34	4.46	0.96	0.01	2.40	35.30	0.95	-0.19
2	3.27	7.17	0.92	0.24	3.06	6.94	0.92	-0.04	3.17	37.10	0.91	-0.22
4	3.79	12.40	0.86	0.84	3.66	11.90	0.84	-0.08	3.72	46.90	0.85	0.02
6	3.93	20.10	0.83	0.87	3.59	17.80	0.82	-0.01	3.57	59.40	0.81	-0.14
8	3.94	28.10	0.83	1.56	3.75	28.10	0.90	0.01	3.53	86.70	0.91	-0.17
12	4.24	108.0	0.75	2.05	3.74	51.10	0.78	-0.02	3.43	173.50	0.89	0.19

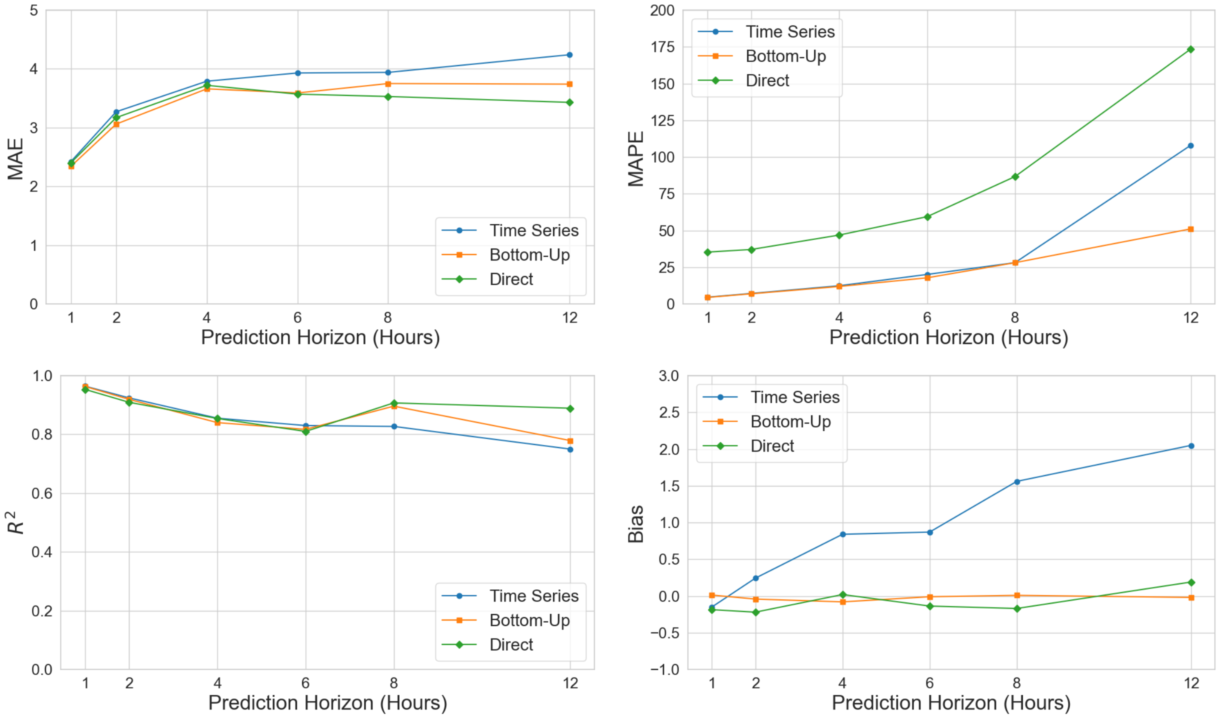


Figure EC.1 Comparison of out-of-sample MAE, MAPE, R^2, and Bias of our three approaches for varying forecasting horizons.

EC.5.2. Generalization to multiple hospitals

In Section 7.2, we extend our analysis to 10 additional EDs in the Seoul metropolitan area.

As presented in Table 10, these ED services differ in their volume, average length-of-stay, and patient severity mix, with larger hospitals associated with longer stays and more severe patients. To appreciate the extent to which longer length of stays are due to different patient severity mix, Table EC.2 presents the average length of stays in these hospitals, per KTAS level. We observe that, indeed, more severe patientsspend more time in the ED. However, larger EDs experience longer LOS than smaller ones, across acuity levels.

Table EC.2 Average Length of Stay (LOS) by KTAS level for each ED from the NEDIS database.

ED	Average LOS by KTAS Level (minutes)				
	KTAS 1	KTAS 2	KTAS 3	KTAS 4	KTAS 5
1	166.9	241.5	192.1	108.6	37.6
2	214.9	242.4	193.4	102.7	47.4
3	175.7	314.6	303.9	156.8	50.1
4	129.2	251.9	229.1	153.5	98.0
5	207.2	307.3	234.4	121.3	64.7
6	291.8	529.8	362.5	148.7	85.4
7	184.8	612.5	517.8	184.9	114.8
8	640.6	624.8	455.5	236.9	175.4
9	677.4	572.8	475.9	257.7	275.3
10	713.0	651.0	542.8	312.9	217.4
Average	509.2	516.0	393.5	205.3	109.0

Figure EC.2 reports a range of accuracy metrics for the three approaches, across the 10 hospitals, for the 4-hour prediction problem. For ease of comparison, Figure EC.3 reports the relative improvement (i.e., reduction) in MAE provided by bottom-up and direct models compared with a times series approach, and Table EC.3 reports the numbers displayed in Figure EC.2.

Table EC.3 Out-of-sample MAE for 4-hour prediction across our 10 EDs.

ED	Mean Absolute Error (MAE)		
	Bottom-Up	Direct	Time Series
1	1.289	1.261	1.460
2	1.102	1.148	1.184
3	1.859	1.832	1.855
4	1.408	1.453	1.512
5	1.508	1.722	1.713
6	3.696	3.088	4.807
7	2.530	2.671	3.343
8	3.912	4.459	5.303
9	4.198	4.316	4.736
10	4.650	4.214	6.643

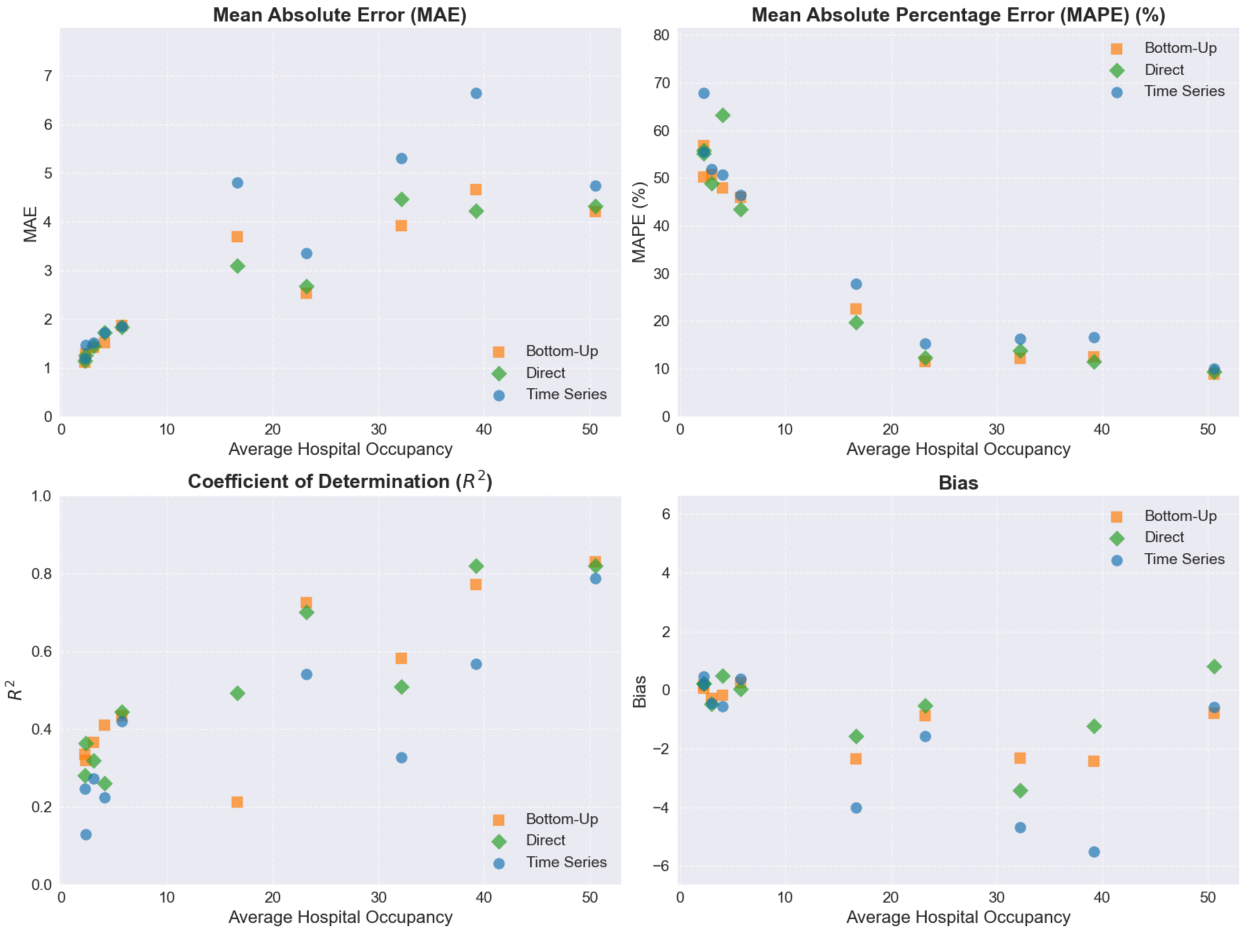


Figure EC.2 **Out-Of-Sample accuracy metrics (MAE, MAPE, R2, Bias) for 4-hour prediction across forecasting methods and different hospitals.**

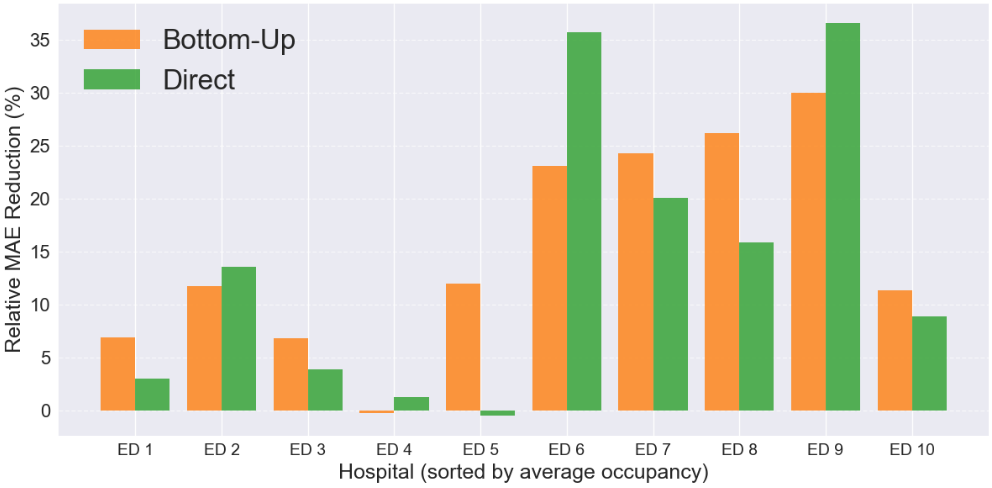


Figure EC.3 **Grouped bar chart showing relative MAE reduction (%) for Bottom-Up and Direct compared to Time Series at 4 hours.**

References

Agarwal N, Biswas B, Singh C, Nair R, Mounica G, H H, Jha AR, Das KM (2021) Early determinants of length of hospital stay: A case control survival analysis among COVID-19 patients admitted in a tertiary healthcare facility of east india. *Journal of Primary Care & Community Health* 12:21501327211054281.

Arora S, W Taylor J, Mak HY (2023) Probabilistic forecasting of patient waiting times in an emergency department. *Manufacturing & Service Operations Management* .

Awad A, Bader-El-Den MB, McNicholas J (2017) Patient length of stay and mortality prediction: A survey. *Health Services Management Research* 30:105 – 120.

Bacchi S, Tan Y, Oakden-Rayner L, Jannes J, Kleinig T, Koblar S (2022) Machine learning in the prediction of medical inpatient length of stay. *Internal Medicine Journal* 52(2):176–185.

Baek H, Cho M, Kim S, Hwang H, Song M, Yoo S (2018) Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. *PLoS One* 13.

Benbelkacem S, Kadri F, Atmani B, Chaabane S (2019) Machine learning for emergency department management. *International Journal of Information Systems in the Service Sector* 11:19–36.

Bertsimas D, Delarue A, Pauphilet J (2024) Simple imputation rules for prediction with missing data: Theoretical guarantees vs. empirical performance. *Transactions on Machine Learning Research* .

Brasel KJ, Lim HJ, Nirula R, Weigelt JA (2007) Length of stay: an appropriate quality measure? *Archives of Surgery* 142(5):461–466.

Breiman L (2001) Random forests. *Machine Learning* 45:5–32.

Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K (2019) Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety* 28(3):231–237, ISSN 2044-5415.

Chang W, Liu Y, Xiao Y, Yuan X, Xu X, Zhang S, Zhou S (2019) A machine-learning-based prediction method for hypertension outcomes based on medical data. *Diagnostics* 9:178.

Chaou CH, Chen HH, Chang SH, Tang P, Pan SL, Yen AMF, Chiu TF (2017) Predicting length of stay among patients discharged from the emergency department—using an accelerated failure time model. *PLoS One* 12:1–11.

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.

Craig CC (1936) On the frequency function of xy. *The Annals of Mathematical Statistics* 7(1):1–15.

Etu EE, Monplaisir L, Arslanturk S, Masoud S, Aguwa C, Markevych I, Miller J (2022) Prediction of length of stay in the emergency department for COVID-19 patients: A machine learning approach. *IEEE Access* 10:42243–42251.

Gill SD, Lane SE, Sheridan M, Ellis E, Smith D, Stella J (2018) Why do 'fast track'patients stay more than four hours in the emergency department? an investigation of factors that predict length of stay. *Emergency Medicine Australasia* 30(5):641–647.

Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning* (MIP Press).

Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. *International Conference on Machine Learning*, 1321–1330 (PMLR).

Haldane J (1942) Moments of the distributions of powers and products of normal variates. *Biometrika* 32(3/4):226–242.

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning* (Springer), 2nd edition.

Kadri F, Abdelkader D, Harrou F, Sun Y (2022) Towards accurate prediction of patient length of stay at emergency department: a gan-driven deep learning framework. *Journal of Ambient Intelligence and Humanized Computing* .

King Z, Farrington J, Utley M, Kung E, Elkhodair S, Harris S, Sekula R, Gillham J, Li K, Crowe S (2022) Machine learning for real-time aggregated prediction of hospital admission for emergency patients. *NPJ Digital Medicine* 5(1):104.

Murphy KP (2023) *Probabilistic Machine Learning: Advanced Topics* (MIT Press).

Naeini MP, Cooper G, Hauskrecht M (2015) Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Stone K, Zwiggelaar R, Jones P, Mac Parthaláin N (2022) A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLoS Digital Health* 1:1–38.

Thomas WJ, Guire KE, Horvat GG (1997) Is patient length of stay related to quality of care? *Journal of Healthcare Management* 42(4):489–507.

Vovk V, Petej I, Nouretdinov I, Ahlberg E, Carlsson L, Gammerman A (2021) Retrain or not retrain: Conformal test martingales for change-point detection. *Conformal and Probabilistic Prediction and Applications*, 191–210 (PMLR).

Wang K, Hussain W, Birge JR, Schreiber MD, Adelman D (2022a) A high-fidelity model to predict length of stay in the neonatal intensive care unit. *INFORMS Journal on Computing* 34(1):183–195.

Wang L, Wang X, Chen A, Jin X, Che H (2020) Prediction of type 2 diabetes risk and its effect evaluation based on the xgboost model. *Healthcare* 8(3).

Wang Z, Liu Y, Wei L, Ji JS, Liu Y, Liu R, Zha Y, Chang X, Zhang L, Liu Q, et al. (2022b) What are the risk factors of hospital length of stay in the novel coronavirus pneumonia (COVID-19) patients? a survival analysis in southwest china. *PLoS One* 17(1):e0261216.

Zadrozny B, Elkan CP (2001) Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *International Conference on Machine Learning*.