Probabilistic Prediction of Arrivals and Hospitalizations in Emergency Departments in Île-de-France

Herbert Susmann, Antoine Chambaz, Julie Josse, Mathias Wargon, Philippe Aegerter, Emmanuel Bacry

Probabilistic Prediction of Arrivals and Hospitalizations in Emergency Departments in Île-de-France

Herbert Susmann[1,*], Antoine Chambaz[2,3], Julie Josse[4], Mathias Wargon[5, 6], Philippe Aegerter[7,8,9], and Emmanuel Bacry[1]

[1]CEREMADE (UMR 7534), Université Paris-Dauphine PSL, Place du Maréchal de Lattre de Tassigny, Paris, 75016, France
[2]Université Paris Cité, CNRS, MAP5, F-75006 Paris, France
[3]Fédération Parisienne de Modélisation Mathématique, CNRS FR 2036
[4]Inria PreMeDICaL team, Idesp, Université de Montpellier
[5]Paris Area Emergency and Unscheduled Care Regional Observatory, Saint-Denis, France
[6]Emergency Department, Saint-Denis Hospital, Saint-Denis, France
[7]Epidemiology and Public Health Service, AP-HP, Hôpitaux Universitaires Paris-Saclay, Boulogne, France
[8]University of Versailles Saint-Quentin, Versailles, France
[9]INSERM CESP U1018, Université Paris-Saclay, Le Kremlin-Bicêtre, France
[*]Corresponding author: Herbert Susmann, herbps10@gmail.com

April 9, 2024

Abstract

Background Forecasts of future demand is foundational for effective resource allocation in emergency departments (EDs). As ED demand is inherently variable, it is important for forecasts to characterize the range of possible future demand. However, extant research focuses primarily on producing point forecasts using a wide variety of prediction algorithms. In this study, our objective is to generate point and interval predictions that accurately characterize the variability in ED demand using ensemble methods that combine predictions from multiple base algorithms based on their empirical performance.

Methods Data consisted in daily arrivals and subsequent hospitalizations at 72 emergency departments in Île-de-France from 2014–2018. Additional explanatory variables were collected including public and school holidays, meteorological variables, and public health trends. One-day ahead point and 80% interval predictions of arrivals and hospitalizations were produced by predicting the 10%, 50%, and 90% quantiles of the forecast distribution. Quantile prediction algorithms included methods such as ARIMAX, variations of random forests, and generalized additive models. Ensemble predictions were then formed using Exponentially Weighted Averaging, Bernstein Online Aggregation, and Super Learning. Prediction intervals were post-processed using Adaptive Conformal Inference techniques. Point predictions were evaluated by their Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), and 80% interval predictions by their empirical coverage and mean interval width.

Results For point forecasts, ensemble methods achieved lower average MAE and MAPE than any of the base algorithms. All of the base algorithms and ensemble methods yielded prediction intervals with near optimal empirical coverage after conformalization. For hospitalizations, the shortest mean interval widths were achieved by the ensemble methods.

Conclusions Ensemble methods yield joint point and prediction intervals that adapt to individual EDs and achieve better performance than individual algorithms. Conformal inference techniques improves the performance of the prediction intervals.

Keywords Emergency department, Time series forecasting, Machine learning, Ensemble learning, Conformal inference

1 Introduction

Demand for emergency department (ED) services has increased in many countries worldwide over the long-term (Pines et al., 2011; Grall, 2015; Colineaux et al., 2019). ED crowding occurs when demand surpasses available resources, and has well-established negative effects including worse patient health outcomes and reduced adherence by practitioners to treatment guidelines (Bond et al., 2007; Hoot and Aronsky, 2008; Moskop et al., 2009; Gacki-Smith et al., 2009; Morley et al., 2018). Because many patients arriving to EDs require subsequent hospitalization, increased ED demand may lead to higher demand for hospital beds. If there are not sufficiently many hospital beds available then patients must wait in the ED until being admitted (a status referred to as "boarding"). Boarders may not receive the same quality of care as they would after admission and may experience worse health outcomes (Liu et al., 2011; Roussel et al., 2023). In addition, inability to admit patients in a timely fashion contributes itself to ED crowding by reducing outflow.

From the point of view of a health system, ED crowding and long boarding times can be ameliorated by increasing available resources, better managing existing resources, or reducing demand for ED services. One possible tool for improving resource management and planning is forecasting ED demand (Eitel et al., 2010). As ED crowding is related both to inflow (patients arriving) and outflow (discharge, hospitalization, or death), it is useful to have forecasting of both arrivals (inflow) and hospitalizations (outflow). Allocating resources sufficient for forecasted average ED demand will inevitably lead to bottlenecks during peaks driven by natural variability in ED arrivals (Higginson et al., 2011). Thus, forecasts that accurately capture the expected variability in demand are needed for planning.

A challenge faced by practitioners seeking to design an ED demand forecasting system is choosing between the wide variety of statistical and machine learning algorithms that have been previously applied to the problem. Algorithms described in the literature are based on classical time series models such as ARIMA (Choudhury and Urena, 2020), regression models (Marcilio et al., 2013), machine learning algorithms such as support vector machines (Zlotnik et al., 2015), and neural networks (Zhao et al., 2022), to name only a few examples. Choosing between the numerous available prediction algorithms is difficult because it is almost never known a-priori which algorithm will perform best in a new setting. In addition, it is possible that the best algorithm will vary for different EDs or that the best-performing algorithm will change over time as data accumulates. As such, ensemble methods that combine predictions from multiple algorithms, adaptively favoring those that perform the best, are desirable, as they obviate the need to choose and rely on an algorithm in advance (van der Laan et al., 2007). There has been significant research in statistics and computer science analyzing the properties of various methods of forming ensembles; in general, under mild assumptions ensemble are expected to perform as well as the best input algorithms (Cesa-Bianchi and Lugosi, 2006; Benkeser et al., 2018; Wintenberger, 2017). Empirically, ensembles have been found to perform well for point forecasts: for example, ensemble methods performed consistently better than relying on a single method for predicting patient demand at Urgent Care Clinics in New Zealand (Maddigan and Susnjak, 2022).

The variability inherent to ED demand can be addressed via methods that produce probablistic forecasts, such as in the form of prediction intervals. Defining quantiles of the predictive distribution of arrivals or hospitalizations as the statistical parameter of interest is therefore natural: the 50% quantile (median) can be used as point prediction, and upper and lower quantiles can be used to form prediction intervals. In this work, we focus on producing 80% prediction intervals formed from estimates of the 10% and 90% conditional quantiles. This approach draws on the long line of research into conditional quantile estimation, for which numerous statistical and machine learning approaches have been developed.

Using ensemble methods to predict the quantiles used to form prediction intervals alleviates the need to choose one particular prediction algorithm for every ED, as we instead adaptively choose the best-performing

algorithms. However, there is no guarantee that the prediction intervals produced by the ensemble will have good finite sample properties (similarly, there is no guarantee that any of the base methods taken alone will yield accurate prediction intervals). For example, we would hope that the 80% prediction intervals produced by ensemble quantile predictions will indeed include the observed number of arrivals (or hospitalizations) nearly 80% of the time. In the statistics and machine learning communities there has been a surge of interest in conformal inference, a family of techniques for forming valid prediction intervals based on the output of any prediction method (Vovk et al., 2005). Traditional conformal inference techniques hinge primarily on the assumption of exchangeability of the observed data (Angelopoulos and Bates, 2023). As this assumption does not typically hold for time series data, recent research has adapted conformal inference techniques this setting. In particular, Adaptive Conformal Inference (ACI) is a family of algorithms that adjust prediction intervals in response to the observations (Gibbs and Candès, 2021, 2022; Zaffran et al., 2022). For example, if the prediction intervals from an ensemble are systematically too small or too large, this will be corrected using ACI. We investigate the use of ACI to improve the empirical performance of the prediction intervals in our dataset.

In this work, we describe an integrated forecasting pipeline for point and prediction intervals of ED arrivals and subsequent hospitalizations. The pipeline incorporates three major components: in the first step, a library of base algorithms are trained and used to produce initial quantile predictions. In the second step, the initial predictions for each quantile are weighted according to their prior empirical performance to form an ensemble prediction. In the final step, the ensemble prediction intervals are post-processed using ACI to improve their performance in finite horizons. We apply the pipeline to generate point forecasts and 80% prediction intervals for arrivals and hospitalizations at EDs in a regional health network.

The rest of the paper unfolds as follows. Section 2 introduces the data and a descriptive analysis of arrivals and hospitalizations. Section 3 formalizes the goal of producing point and prediction intervals as a conditional quantile estimation task and defines evaluation metrics. Section 4 describes in detail each step of the integrated forecasting pipeline. Section 5 presents empirical results based on applying the pipeline to forecast demand in EDs in Île-de-France, France. Section 6 discusses our findings and avenues for future research.

Prior Work ED demand forecasting has been the subject of significant research. Existing approaches vary widely in terms of the forecasting goal (such as focusing on daily, weekly, or monthly forecast horizons), the statistical and machine learning algorithms that are applied, the external covariates used, and how the results are evaluated. We refer to several systematic reviews for a thorough survey (Wargon et al., 2009; Gul and Celik, 2020; Jiang et al., 2022). Most papers forecast the number of ED arrivals; relatively fewer papers attempt to predict in advance the number of hospitalizations originating from an ED, and are typically applied to a relatively small number of EDs (Jiang et al., 2022). Methodologically, our approach is closely related to that of Rostami-Tabar et al. (2023), who apply several quantile estimation algorithms to produce predictions of hourly emergency department arrivals at a large emergency department. Our work can be thought of as complementary to their approach in that we also focus on probabilistic forecasting via quantile prediction, but we take the additional step of aggregating predictions from many algorithms using ensemble methods. We also extend their approach by post-processing the prediction intervals using conformal inference techniques to improve finite sample performance.

2 Descriptive Analysis

2.1 Data

The network of EDs in the Île-de-France region of France, encompassing the Paris metropolitan area, serves a population of over 12 million inhabitants. EDs in the network are located within health establishments (such as hospitals). In some cases, a health establishment may house multiple EDs; for example, a hospital may separate their pediatric and adult emergency services. All the EDs transmit summary arrival data daily to a central repository managed by the Regional Health Agency (*Agence Régionale de Santé*). The raw dataset for this study is derived from this source and includes visits from 2014-2018.

Each arrival in the raw dataset has metadata indicating the ED, date and time of arrival, date and time of discharge, mode of discharge, and diagnosis code (ICD-10). The times of arrival and discharge

have been previously anonymized by the addition of random noise (±2m30s). Arrivals with discharge coded as hospitalization in the same health establishment or as hospitalization in another establishment were categorized as hospitalizations. The aggregate number of arrivals and hospitalizations per ED by day of arrival was then calculated. When no arrivals are reported for an ED during a day, it is impossible to know whether the ED indeed received no arrivals or if the data transfer failed for that day. We treat any such day as having missing data for arrivals (missing data will be discussed in the paragraph below). To address the possibility that EDs exceptionally closed for part of the day, we calculated for each ED the mean number of hours per day in which there was at least one arrival. Days with less than one-half the mean number of hours with at least one arrival were flagged as probable days in which the ED was exceptionally closed, and were removed from the dataset. The analysis dataset was subsequently formed by selecting the emergency departments that reported at least 200 days of data and at least one hospitalization in each of the years 2014–2018, for a total of 72 EDs.

Data Quality Arrival data may be missing or inaccurate for several reasons. Data related to every arrival were transmitted once per day from each emergency department to a centralized data repository. Data for some days are either not available in the analysis dataset or the number of arrivals or hospitalizations is reported as zero due to technical failures in the transmission process. There may be outliers in the dataset that arise for administrative reasons. For example, some emergency departments may be exceptionally closed for a day or part of a day, resulting in a smaller than usual number of arrivals and hospitalizations. Data that are successfully reported may be inaccurate due to human error in the data collection process. More accurate data collection would be inherently very difficult in this setting given the exigencies of emergency medicine. The goal of our work is therefore to predict the observed arrivals and hospitalizations, which we emphasize is not equivalent to predicting the true number of arrivals and hospitalizations which may have been different.

Enrichment Several additional variables thought to be possibly related to ED demand were collected. First, French national holidays and public school vacation periods were obtained (data.gouv.fr, 2023; Augusti, 2023). In addition, an indicator was added for days following a national holiday. An indicator was also added for any Friday immediately following a national holiday, as they are commonly taken as a vacation day. Indicators of the dates January 1st, July 14th, December 25th, and December 31st were also included as covariates to capture any pattern relating to the corresponding holidays. Meteorological variables (daily minimum and maximum temperature) were gathered from the closest Automated Surface Observing System location to each ED with data availability covering the study period (Salmon, 2023). Missing daily minimum or maximum temperatures were imputed with linear interpolation. Daily precipitation data (measured in tenths of millimeters) were gathered from the nearest monitoring station to each ED in the Global Historical Climatology Network (Menne et al., 2012). One-day lagged values of the meteorological variables were included as covariates, representing the information that would be available at the time of prediction. Weekly incidence of chickenpox, acute diarrhea, and flu-like maladies in the Île-de-France region were sourced the Sentinelles public health surveillance system (Valleron et al., 1986; Flahault et al., 2006). Weather variables were lagged by one day and weekly disease incidence lagged by one week, again to ensure predictions are made based only on covariates available at the time of prediction.

2.2 Descriptive Analysis

To contextualize the goal of point and interval predictions for arrivals and hospitalizations at EDs we performed several descriptive analyses. The emergency departments in the analysis dataset cover a range of activity levels (Table 1). The average daily arrivals at each emergency department ranged from 27.2 to 248.2 (the overall average number of arrivals each day across all emergency departments was 113.1), and the average number of hospitalizations each day ranged from 3 to 49.1 (overall average: 17.3). Temporal trends in arrivals and hospitalizations were heterogeneous across the EDs in the analysis dataset. As an example, Figure 1 displays a selected ED that exhibits immediately apparent seasonal trends in which arrivals and hospitalizations are lower in summer, particularly in August which is a common time for vacation in the region. Some emergency departments exhibited different flows depending on the day of the week; in some cases, arrivals were lower during the weekend and higher on Monday; however, this pattern was not universal.

ED summary measure	Average across all EDs	(range across all EDs)
Most arrivals in one day	193.2	(72.0-474.0)
Average arrivals in one day	113.1	(27.2-248.2)
Fewest arrivals in one day	54.0	(8.0-158.0)
Days of missing arrivals	38.5	(0.0-148.0)
Most hospitalizations in one day	40.5	(10.0-114.0)
Average hospitalizations in one day	17.3	(3.0-49.1)
Fewest hospitalizations in one day	3.5	(0.0-14.0)
Days of missing hospitalizations	38.7	(0.0-148.0)

Table 1: Summary measures of daily arrivals and hospitalizations for EDs in the analysis dataset.

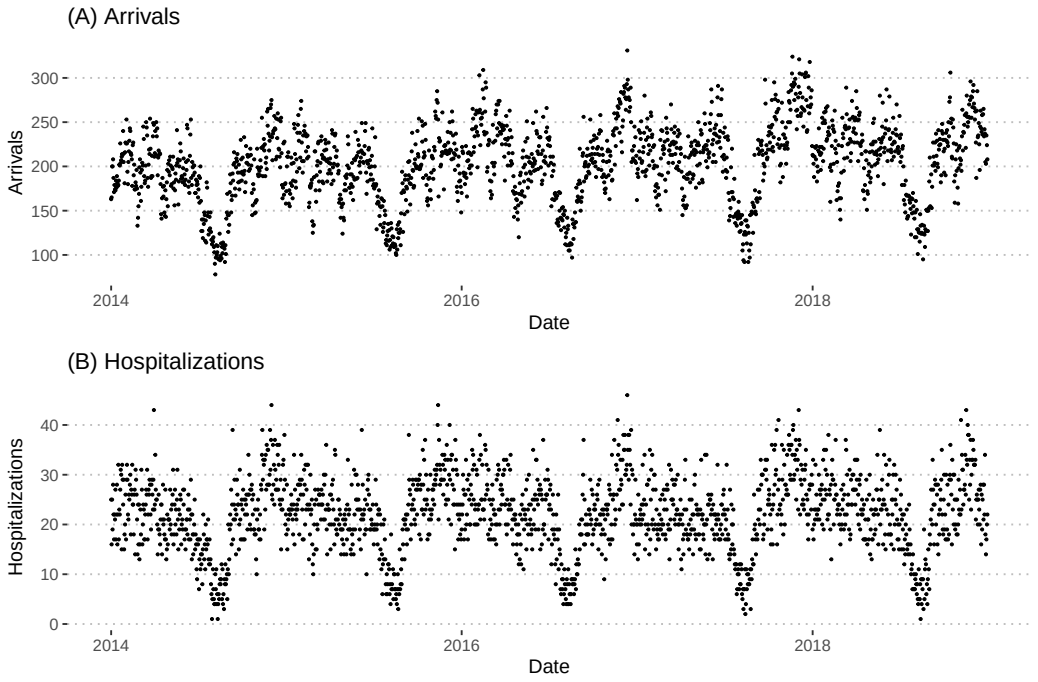

Figure 1: Illustrative time series of daily arrivals (A) and hospitalizations (B) in one ED selected from the analysis dataset.

The raw arrivals dataset includes an ICD-10 diagnosis code which can be used to understand reasons why patients seek emergency treatment. Due to high levels of missingness (25.3% of arrivals are missing a diagnosis code), we only use these data to illustrate broad trends. Hospitalizations for acute bronchiolitis, asthma, and fracture of femur are shown in Figure 2 as an illustrative example. Acute bronchiolitis follows an obvious seasonal trend, peaking in winter, while femur fractures have no discernable pattern; asthma is somewhere in between. These examples illustrate how the number of arrivals and hospitalizations can be seen as the aggregation of many different processes, some of which are more predictable than others. From this point of view it is clear there is a bound on the possible accuracy of point predictions of the daily number of arrivals and hospitalizations. For example, it is unlikely that we will be able to predict exactly how many people will suffer a femur fracture on any particular day. On the other hand, the *variability* in the number of femur fractures is consistent over time. This suggests that it is possible to produce high-quality prediction intervals that accurately characterize the variability in arrivals and hospitalizations.

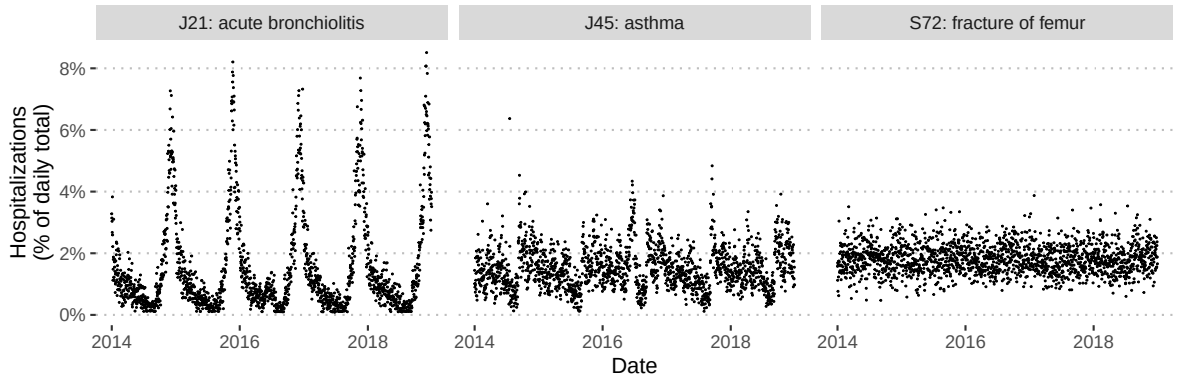

Figure 2: Hospitalizations coded with ICD-10 block code J21 (acute bronchiolitis), J45 (asthma), and S72 (fracture of femur) as percentage of total daily hospitalizations in all EDs in the analysis dataset. Hospitalizations for these three diagnoses show different seasonality, with bronchiolitis following a strong seasonal trend, asthma a weak trend, and fracture of the femur no trend visible to the naked eye.

3 Goals and Metrics

Informally, the overall goal is to make one-day-ahead point and 80% interval predictions for daily arrivals and hospitalizations for each ED in the analysis dataset. First, we formalize this goal statistically as a conditional quantile estimation task. We then discuss the evaluation metrics used to evaluate the quality of the forecasts.

3.1 Forecasting Task

Let Δ_t indicate whether data for the number of arrivals (or subsequent hospitalizations) is available at an ED on day t, where $\Delta_t = 1$ indicates data are available and $\Delta_t = 0$ indicates otherwise. Let y_t be the number of arrivals (or subsequent hospitalizations) at an ED on day t. We seek to predict quantiles of the predictive distribution of y_t conditional on all data available before time day t. Specifically, we define a point prediction $\hat{y}_t$ as the 50% quantile (median) of the predictive distribution. To form a $(1-\beta)\times 100\%$ prediction interval, we predict the $\alpha_1 = \beta/2$-quantile and $\alpha_2 = (1-\beta)/2$-quantile of the predictive distribution, with $\hat{\ell}_t$ the predicted lower quantile and $\hat{u}_t$ the upper quantile. We set $\beta = 0.2$ such that $(1-\beta)\times 100\% = 80\%$ prediction intervals are targeted using estimates of the $\alpha_1 = 10\%$ and $\alpha_2 = 90\%$ quantiles.

3.2 Evaluation Metrics

Evaluating the quality of predictions plays a critical role in our forecasting pipeline. Predictions from a diverse set of base algorithms are evaluated, and the best-performing ones are given more weight in ensemble predictions. Both the point and interval predictions are defined in terms of quantiles of the predictive distribution. As such, we chose to evaluate the estimated quantiles using the quantile loss function (Gneiting, 2011a,b), defined for the α-quantile ($\alpha \in (0,1)$) as

$$L_\alpha(\hat{y}, y) = \begin{cases} \alpha|y-\hat{y}| & \text{if } y \geq \hat{y}, \\ (1-\alpha)|y-\hat{y}| & \text{if } y \leq \hat{y}. \end{cases}$$

where y is the observed value and $\hat{y}$ is the prediction of the α-quantile. Note that when $\alpha = 0.5$, the quantile loss function simplifies to the familiar absolute error loss, which penalizes under- and over-predictions equally. For estimating other quantiles, the quantile loss function can be thought of as an asymmetric version of the absolute error loss. For example, when $\alpha = 0.1$, the loss function penalizes predictions that fall above the observation more than predictions that fall below the observation, thereby encouraging predictions that capture the lower tail of the forecast distribution.

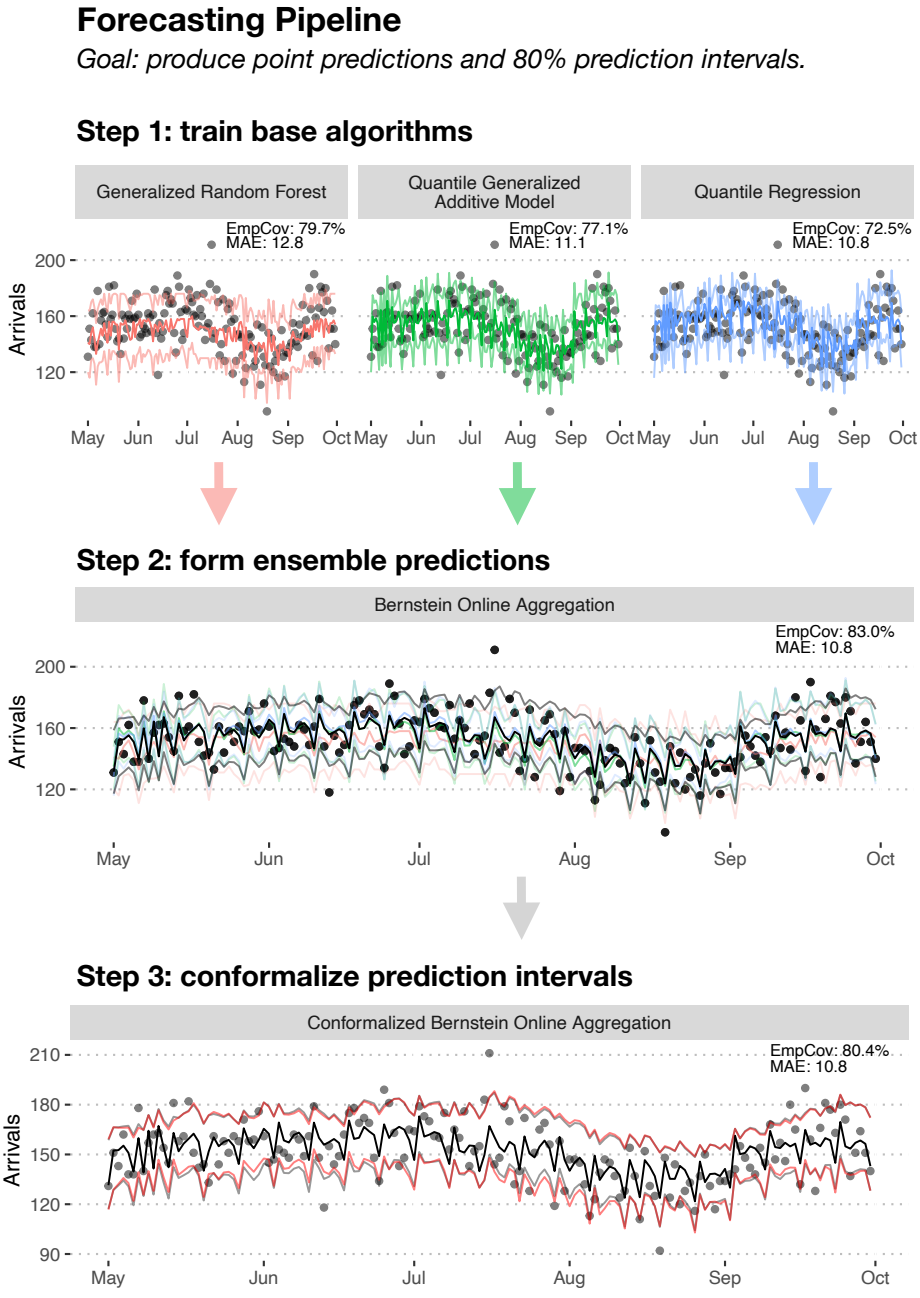

Figure 3: Graphical summary of the forecasting pipeline. In the first step, a set of base algorithms are applied to produce predictions of the 10%, 50%, and 90% quantiles. The 10% and 90% quantiles are used to form 80% prediction intervals. In the second step, an ensemble estimate is formed as weighted combination of the base predictions where well-performing algorithms are given higher weight. In the third step, the prediction intervals are post-processed using Adaptive Conformal Inference techniques to improve their empirical coverage. The empirical coverage (EmpCov, with optimal level 80%) and Mean Absolute Error (MAE, smaller is better) for each method are shown for the ED and time period depicted.

To evaluate a sequence of predictions, we define an empirical risk as the mean of the quantile loss function over all days in which data on the outcome are available at horizon T:

$$R_\alpha(T) = \frac{1}{\sum_{t=1}^{T} \Delta_t} \sum_{t=1}^{T} \Delta_t L_\alpha(\hat{y}_t, y_t).$$

For example, in our evaluations we test the performance of the prediction pipeline for every day in 2018, and thus $T = 365$. Importantly, the α-quantile is a minimizer of the α-quantile empirical risk; we therefore say that the quantile loss function is *consistent* for the α-quantile parameter. This is an important property for a loss function to have because it encourages honest forecasts (Gneiting, 2011a).

As a summary measure for the point predictions $\hat{y}_t$, defined as the median of the forecast distribution, we report the Mean Absolute Error (MAE) at time T:

$$\text{MAE}(T) = 2R_{0.5}(T) = \frac{1}{\sum_{t=1}^{T} \Delta_t} \sum_{t=1}^{n} \Delta_t |y_t - \hat{y}_t|. \tag{1}$$

The MAE is twice the empirical risk of the quantile loss when $\alpha = 50\%$, and therefore the MAE is consistent for the 50% quantile (median) parameter. As a secondary metric we report the Mean Absolute Percentage Error (MAPE), defined at time T as

$$\text{MAPE}(T) = \frac{1}{\sum_{t=1}^{T} \Delta_t} \sum_{t=1}^{T} \frac{\Delta_t |y_t - \hat{y}_t|}{\max\{y_t, 1\}}. \tag{2}$$

The MAPE is a commonly used metric in ED demand forecasting studies; however, it has well-documented drawbacks as an error metric for point predictions (Armstrong and Collopy, 1992; Makridakis, 1993; McKenzie, 2011). Specifically, it is non-symmetric: overpredictions and underpredictions incur different penalties under the MAPE. While asymmetry is desirable if we wish to estimate a lower or upper quantile, for example, it leads to unintended consequences when used as an error metric of a point prediction intended to capture the central tendency of a forecast distribution. In particular, minimizing the MAPE will encourage predictions that systematically underestimate the observations. As such, from a statistical point of view we do not recommend using the MAPE, but we report it to facilitate comparisons with other published approaches.

To evaluate the quality of a sequence of prediction intervals we report the empirical coverage, defined at time T as:

$$\text{EmpCov}(T) = \frac{1}{\sum_{t=1}^{T} \Delta_t} \sum_{t=1}^{T} \Delta_t \mathbb{I}[\hat{\ell}_t < y_t < \hat{u}_t]. \tag{3}$$

The empirical coverage is the proportion of observations (arrivals or hospitalizations) that fell within their respective prediction intervals. In addition, we also report the mean width of the prediction intervals, defined at time T as

$$\text{MeanWidth}(T) = \frac{1}{\sum_{t=1}^{T} \Delta_t} \sum_{t=1}^{T} \Delta_t (\hat{u}_t - \hat{\ell}_t). \tag{4}$$

In general, we seek *sharp* prediction intervals that have small mean width while maintaining optimal empirical coverage.

4 Forecasting Pipeline

In this section we describe an integrated forecasting pipeline for point and interval predictions of daily arrivals and hospitalizations. The pipeline begins by training a set of base algorithms (Step 1) which are then weighted according to their empirical performance and combined into ensemble predictions (Step 2). The resulting prediction intervals are then post-processed using conformal inference techniques to improve their finite-sample performance (Step 3). The pipeline is summarized visually in Figure 3.

Algorithm	R package	Citations
ARIMA, ARIMAX	`forecast`	Hyndman and Khandakar (2008); Hyndman et al. (2023)
Quantile Forests	`grf`	Athey et al. (2019)
Distributional Random Forests	`drf`	Ćevid et al. (2022)
Gradient Boosted Machine	`gbm`	Greenwell et al. (2022)
Quantile Generalized Additive Models	`qgam`	Fasiolo et al. (2020, 2021a)
Quantile Regression	`quantreg`	Koenker (2005); Koenker et al. (2017)

Table 2: Base algorithms used for point and prediction interval estimation.

4.1 Step 1: train base algorithms

The first step in the forecasting pipeline is to apply a variety of algorithms to predict one-day ahead ED arrivals and hospitalizations. We build a diverse library of algorithms encompassing a variety of approaches, as we do not know a-priori which algorithms will perform best. The library we used included classical statistical time-series models (ARIMA, ARIMAX), regression approaches (quantile regression, quantile generalized additive models) as well as machine learning algorithms (quantile forests, distributional random forests, generalized boosted regression models). All analyses were conducted using R version 4.2 (R Core Team, 2022). A summary of the base algorithms is given in Table 2. An overview of the base algorithms and implementation details, including choices of tuning parameters, is available in the appendix.

As point prediction benchmarks we used two naive algorithms based on predicting previous observed values. First, we predict that the outcome will be equal to the outcome on the previous day (formallly, $\hat{y}_t = y_{t-1}$). To capture the possibility of day of the week effects, we also tested a naive algorithm that predicts the outcome from one week before ($\hat{y}_t = y_{t-7}$).

As an alternate benchmark we adapted an algorithm used by French hospitals to predict minimal hospital bed requirements, referred to as *Besoin Journalier Minimal en Lits*, (BJML; Daily Minimal Bed Requirement). The BJML forecast for day a t falling in week $w \in \{1, \ldots, 52\}$ is defined as the empirical 25% quantile of historical hospitalizations in an ED for days falling in the same week w. BJML is thus designed to capture basic seasonal patterns in hospitalizations. We use a straightforward generalization of the BJML to forecast any α-quantile of ED arrivals and subsequent hospitalizations.

4.2 Step 2: form ensemble predictions

In the second step of the pipeline we combine the predictions from each of the base algorithms as a weighted average according to the algorithm's past performance. More formally, suppose we have $k = 1, \ldots, K$ base algorithms, each yielding a prediction $\hat{y}^{\alpha}_{t,k}$ of the α-quantile of the outcome on day t. We will consider ensembles formed as a weighted combination of the predictions from each of the base algorithms:

$$\hat{y}_t = \sum_{k=1}^{K} w_{t,k} \hat{y}_{t,k},$$

where $w_{t,k} > 0$ for $k \in 1, \ldots, K$ and $\sum_{k=1}^{K} w_{t,k} = 1$ (that is, $\mathbf{w}_t = \{w_{t,1}, \ldots, w_{t,K}\}$ is in K-dimensional simplex, denoted $\mathcal{W}$). The goal is to choose weights that yield an ensemble prediction that performs well in terms of empirical risk. There has been a vast amount of research on the subject with many different algorithms for choosing the weights proposed (Cesa-Bianchi and Lugosi, 2006). We compared three different common approaches, described briefly below. In each case, the ensemble algorithm is applied separately to form estimates of the 10%, 50%, and 90% quantiles of the forecast distribution of the outcome (arrivals or hospitalizations). The 10% and 90% quantile estimates are then used to form 80% prediction intervals.

Super Learning In the Super Learning approach (van der Laan et al., 2007; Benkeser et al., 2018; Ecoto et al., 2021), the weights are chosen to minimize the empirical risk of the ensemble in hindsight: at any time $t < T$,

$$w_{t+1} = \arg\min_{w \in \mathcal{W}} R_{\alpha}(t)$$

This approach is intuitive: each day, we use the weights that would have yielded the best performance had they been used in every previous day. This algorithm is also known as *Follow the leader* in computer science as the best-performing base algorithm receives the highest weight (Cesa-Bianchi and Lugosi, 2006). A benefit of Super Learning is that it requires no tuning parameters. However, finding the weights requires solving an optimization problem, which can be computationally demanding.

Exponentially Weighted Average The Exponentially Weighted Average (EWA) algorithm (Cesa-Bianchi and Lugosi, 2006) updates the weights assigned to each of the candidate algorithms at time t according to the rule

$$w_{k,t+1} = \frac{w_{k,t} \exp\left(-\gamma L_\alpha(\hat{y}_{k,t}, y_t)\right)}{\sum_{j=1}^{K} w_{j,t} \exp\left(-\gamma L_\alpha(\hat{y}_{j,t}, y_t)\right)},$$

where $\gamma > 0$ is a learning rate parameter that controls how quickly the weights can change as new information is accumulated about the performance of the candidate algorithms. We use the implementation of EWA in the `opera` `R` package that aggregates over a grid of learning rates based on their performance (Gaillard et al., 2023). The weights at the first timestep can be chosen arbitrarily.

Bernstein Online Aggregation The Bernstein Online Aggregation (BOA) algorithm is similar to EWA, but features stronger theoretical guarantees (Wintenberger, 2017). BOA updates the weights at time t according to the modified rule

$$w_{k,t+1} = \frac{w_{k,t} \exp(-\gamma L_\alpha(\hat{y}_{k,t}, y_t)(1 + \gamma L_\alpha(\hat{y}_{k,t}, y_t)))}{\sum_{j=1}^{K} w_{k,t} \exp(-\gamma L_\alpha(\hat{y}_{k,t}, y_t)(1 + \gamma L_\alpha(\hat{y}_{k,t}, y_t)))},$$

where $\gamma > 0$ is again learning rate parameter. We use a version of BOA as implemented in the `opera` package that calibrates the learning rate automatically (Wintenberger, 2017; Gaillard et al., 2023).

4.3 Step 3: conformalize prediction intervals

Our approach so far for forming $(1 - \beta) \times 100\%$ prediction intervals hinges on being able to well estimate lower $(\beta/2)$ and upper $(1 - \beta/2)$ quantiles. We would expect the prediction intervals formed from estimates of these quantiles to include the observed data nearly $(1-\beta)\%$ of the time. However, in practice it is possible for the prediction intervals to fail to achieve optimal coverage. As such, we turn to recent developments in conformal inference techniques to post-process the prediction intervals. In general, the goal of conformal inference methods is to produce prediction intervals that provably achieve the optimal level of coverage without making any distributional assumptions on the underlying data generating process and using the predictions of any algorithm as input (Shafer and Vovk, 2008; Angelopoulos and Bates, 2023). We will draw on a recent line of research on ACI methods which were developed specifically for generating online prediction intervals for time series (Gibbs and Candès, 2021, 2022; Zaffran et al., 2022). The idea of these methods is to adaptively adjust the prediction intervals generated by a prediction method in response to the observed data. In a nutshell, if the estimated prediction intervals are too narrow and are not covering the observed data then they are adjusted to be slightly wider, and vice versa if the intervals are too wide.

More formally, suppose we have a prediction interval with lower and upper bounds $\hat{\ell}_t$ and $\hat{u}_t$ for an ED on day t. We adjust the size of the prediction interval according to a parameter $\theta_t \in \mathbb{R}$: that is, the adjusted prediction interval has lower bound $\hat{\ell}_t - \theta_t$ and upper bound $\hat{u}_t + \theta_t$. In the original ACI algorithm of Gibbs and Candès (2021), after each day we set the parameter value θ_{t+1} for the following day according to a simple rule, depending on whether number of arrivals (hospitalizations) on day t was inside or outside of the prediction interval:

- If the observed number of arrivals (hospitalizations) on day t fell within the prediction interval, then the interval on the next day is made shorter: $\theta_{t+1} = \theta_t - \gamma\alpha$.
- If the observed number of arrivals (hospitalizations) fell outside the prediction interval on day t, then the interval on the next day is made larger: $\theta_{t+1} = \theta_t + \gamma(1 - \alpha)$.

While the simplicity of this approach is appealing, in practice it is difficult to use because it is not clear how to set the learning rate parameter γ. To solve this problem, the Aggregated ACI (AgACI) method builds an ensemble of base ACI algorithms, each using a different learning rate (Zaffran et al., 2022). The lower and upper bounds of the resulting prediction intervals are then combined using an overarching online aggregation of experts algorithm. Following Zaffran et al. (2022) we use Bernstein Online Aggregation (Wintenberger, 2017) as the aggregation method. Although there have been several other ACI algorithms proposed, in practical settings they have been found to yield similar prediction intervals; therefore we consider only AgACI in this work for simplicity Susmann et al. (2023). We used the implementation of AgACI available in the `ConformalInference` `R` package Susmann et al. (2023).

5 Results

We performed a temporal cross-validation exercise to investigate the performance of our methods. The forecasting pipeline was used to predict arrivals and hospitalizations for each ED and each day in the analysis dataset covering 2018. The base algorithms were trained separately on each ED in the analysis dataset, and were retrained after each month in 2018. After each day, the ensemble weights and conformal prediction intervals were updated using the true number of arrivals and hospitalizations observed on that day. For the point forecasts we report the Mean Absolute Error (1) and Mean Absolute Percentage Error (2) over days in 2018. For prediction intervals we report the empirical coverage (3) and mean interval width (4) over 2018. Illustrative point and interval predictions for one ED from May to October 2018 are shown in Figure 3. Additional illustrative predictions based on Bernstein Online Aggregation for one ED in the analysis dataset for all of 2018 are shown in Appendix Figure 6.

5.1 Point forecasts

The performance of the point forecasts for arrivals and hospitalizations from each of the benchmarks, base algorithms and ensemble methods, averaged across all the EDs in the analysis dataset, is shown in Table 3. The benchmark methods, which include BJML (the median of arrivals or hospitalizations from the same week in previous years) and predicting with the number of arrivals or hospitalizations observed one day or seven days before the target date (referred to as $\hat{y}_{t-1}$ and $\hat{y}_{t-7}$, respectively), had higher MAE and MAPE than any of the base algorithms. Among the base algorithms, quantile regression achieved the lowest average MAE for arrivals and tied for the lowest average MAE for hospitalizations (10.1 and 3.5, respectively). The ensemble methods constructed using Exponentially Weighted Averaging, Super Learning, and Bernstein Online Aggregation generally exhibited as good or better performance than the base algorithms. For arrivals the forecasts constructed using Bernstein Online Aggregation achieved the lowest average MAE (10.0), and for hospitalizations Bernstein Online Aggregation, Exponentially Weighted Averaging, and Super Learner tied for the lowest average MAE (3.4). While the ensemble methods were not constructed to minimize the average MAPE, Bernstein Online Aggregation had the lowest (or tied for lowest) average MAPE of all the algorithms considered.

The best-performing base algorithm depended on the ED and on whether arrivals or hospitalizations were the forecast target. Table 4 shows the percentage of EDs for which each of the base algorithms achieved the lowest MAE or MAPE for arrivals and hospitalizations. For arrivals, quantile regression was the best-performing algorithm in 69.4% of EDs in terms of MAE, followed by Quantile Generalized Additive Model (best MAE in 19.4% of EDs) and Gradient Boosted Machine (best MAE in 6.9% of EDs). For hospitalizations, Quantile Regression a Quantile Generalized Additive Model were tied, each having the best MAE in 25.0% of EDs, followed by Gradient Boosted Machines, Distributinoal Random Forest, and ARIMA (best MAE in 8.3% of EDs). The best-performing methods were similar in terms of the MAPE. That none of the base algorithms dominated across all EDs for either arrivals or hospitalizations suggests that the use of ensemble methods, which adaptively upweight the best-performing method adaptively, is warranted.

To further illustrate the variability of algorithm performance across EDs, Figure 4 shows the MAE for arrivals and hospitalizations by ED for Bernstein Online Aggregation, Gradient Boosted Machine, and Quantile Regression. As suggested previously in Table 4, there was not a consistent trend across EDs in terms of Gradient Boosted Machine or Quantile Regression having lower MAE. The ensemble Bernstein Online Aggregation method tended to match or outperform the performance of the base algorithms. This

	Arrivals		Hospitalizations	
Algorithm	MAE	MAPE	MAE	MAPE
Benchmark				
BJML	15.2	13.1%	4.1	40.9%
Y_{t-1}	15.1	14.1%	4.9	43.2%
Y_{t-7}	14.5	13.8%	4.7	41.9%
Base Algorithms				
ARIMA	11.8	11.2%	3.6	34.3%
ARIMAX	11.7	11.1%	3.7	34.6%
Distributional Random Forest	11.6	10.6%	3.6	31.8%
Generalized Random Forest	11.3	10.5%	3.5	31.6%
Gradient Boosted Machine	10.5	10.0%	3.5	32.6%
Quantile Generalized Additive Model	10.3	9.9%	3.5	33.0%
Quantile Regression	10.1	9.7%	3.5	32.2%
Ensemble Methods				
Bernstein Online Aggregation	**10.0**	**9.6%**	**3.4**	**31.2%**
Exponentially Weighted Average	10.1	**9.6%**	**3.4**	31.7%
Super Learner	10.1	9.7%	**3.4**	31.4%

Table 3: Performance of point predictions from each of the base algorithms and ensemble methods for arrivals and hospitalizations in terms of Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) averaged over all EDs in the analysis dataset. The best-performing algorithm(s) in terms of each metric are bolded.

suggests that the ensemble was able to detect which methods perform better in a particular ED and give their predictions more weight in the combined prediction.

5.2 Prediction intervals

The performance of the predictions intervals before and after conformalization for arrivals and hospitalizations is shown in Table 5. The benchmark method BJML, which forms interval predictions by using empirical quantiles of arrivals and hospitalizations in the same week of previous years, had empirical coverage farthest from the optimal 80% level before conformalization. Among the base algorithms, empirical coverage varied widely. The classical time-series methods ARIMA and ARIMAX (ARIMA with covariates) had near-optimal empirical coverage for both arrivals and hospitalizations before conformalization, although their mean interval widths were larger than other methods. Finally, the ensemble methods yielded prediction intervals that slightly undercovered, although the intervals from Bernstein Online Aggregation had near optimal 79.3% empirical coverage for arrivals and hospitalizations, respectively, before conformalization.

After conformalization, prediction intervals from the benchmark BJML method, the base algorithms, and the ensemble methods all achieved near optimal empirical coverage ranging from 79.6% to 80.5% for arrivals and 80.3% to 81.4% for hospitalizations. The mean interval widths varied by algorithm even after conformalization, illustrating that even though all methods achieved near optimal empirical coverage, the sharpness of the intervals varied across methods. For hospitalizations, the ensemble methods had lower mean interval widths than any of the base algorithms, and specifically Bernstein Online Aggregation had the smallest mean interval widths (11.7).

To understand how performance of the intervals from the ensemble methods varied across EDs, Figure 5 shows the empirical coverage and mean interval widths for every ED before and after conformalization, illustrating the ameliorative effect of conformalization. In general, the original prediction intervals from the ensemble methods tended to systematically undercover, which was corrected by conformalization. Even for Bernstein Online Aggregation, which yielded intervals with near optimal empirical coverage before conformalization in aggregate across all EDs, conformalization improves the performance of the intervals by correcting EDs with intervals that tended to under- or over-cover. Appendix Figures 7 and 8 show similar results for the MAPE.

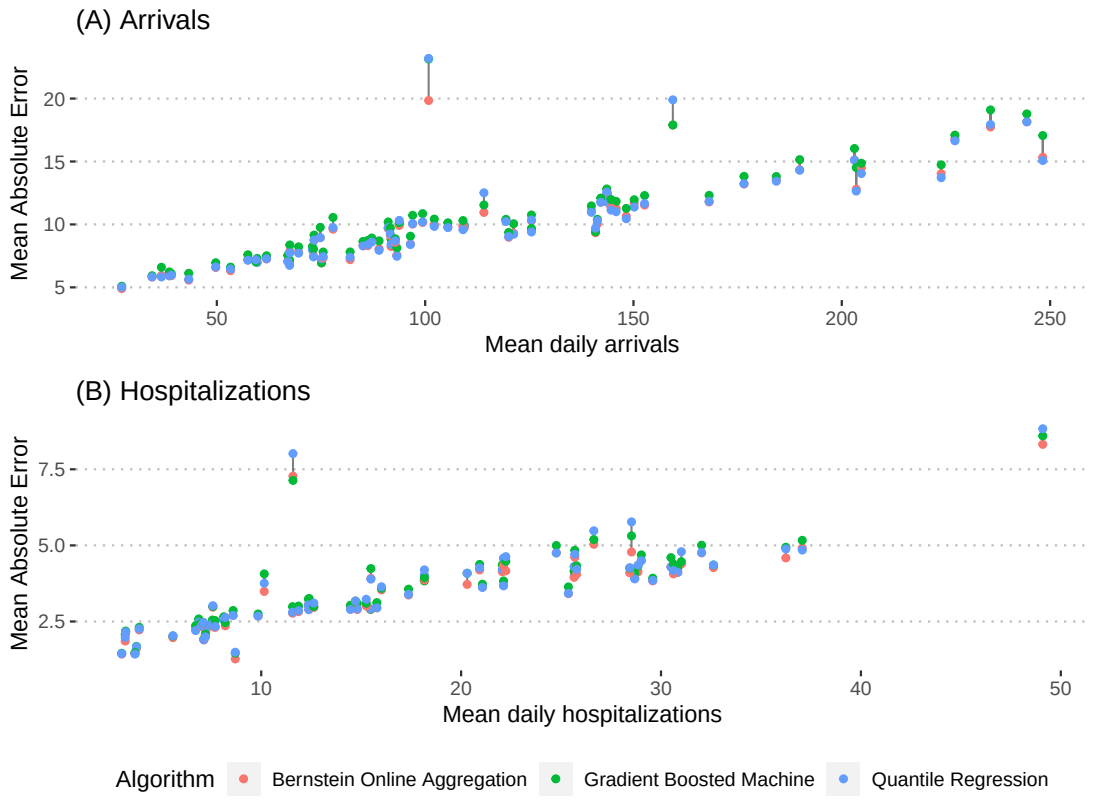

Figure 4: Mean absolute error of point predictions for (A) arrivals and (B) hospitalizations from Bernstein Online Aggregation, Gradient Boosted Machine, and Quantile Regression. Each point shows the performance of an algorithm in one ED, and points from the same ED are connected by a line. The x-axis is the mean daily arrivals and hospitalizations across all days in the analysis dataset. Between Gradient Boosted Machines and Quantile Regression, the better performing method depends on the ED. The ensemble method Bernstein Online Aggregation tended to achieve the lowest MAE across all EDs.

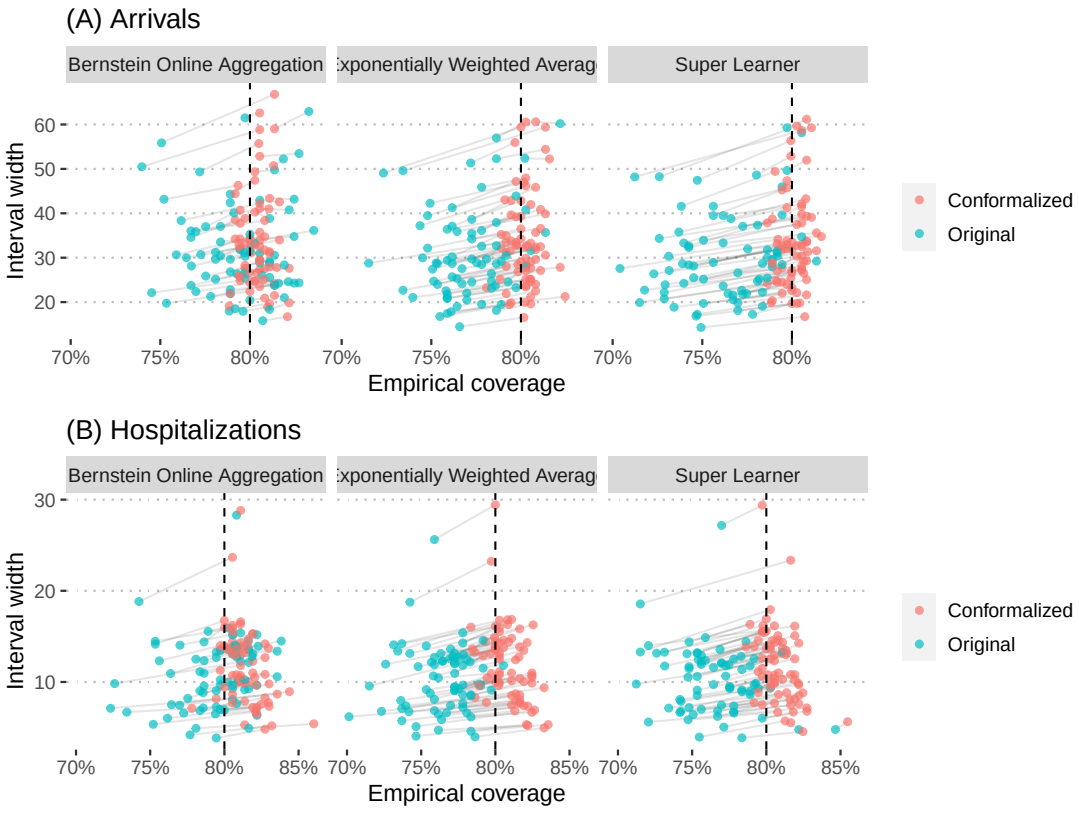

Figure 5: Mean interval width and empirical coverage of ensemble prediction intervals for each ED in the analysis dataset before and after conformalization. For each of the ensemble algorithms, the empirical coverage is closer to the optimal 80% level after conformalization.

Base Algorithm	Achieved best MAE in X% of EDs	Achieved best MAPE in X% of EDs
Arrivals		
ARIMA	0.0%	0.0%
ARIMAX	2.8%	0.0%
Distributional Random Forest	1.4%	2.8%
Gradient Boosted Machine	6.9%	13.9%
Generalized Random Forest	0.0%	2.8%
Quantile Generalized Additive Model	19.4%	13.9%
Quantile Regression	**69.4%**	**66.7%**
Hospitalizations		
ARIMA	8.3%	5.6%
ARIMAX	1.4%	2.8%
Distributional Random Forest	8.3%	15.3%
Gradient Boosted Machine	8.3%	9.7%
Generalized Random Forest	**25.0%**	**23.6%**
Quantile Generalized Additive Model	23.6%	19.4%
Quantile Regression	**25.0%**	**23.6%**

Table 4: Percentage of EDs for which each of the base algorithms achieved the lowest Mean Absolute Error (MAE) or Mean Absolute Percentage Error (MAPE) among all of the base algorithms. The base algorithms that performed best in terms of MAE and MAPE in the highest percentage of EDs is bolded. No single base algorithm dominated the others in terms of having the lowest MAE or MAPE across all EDs.

6 Discussion

A core challenge in ED forecasting is due to the inherent variability in demand for ED services. The descriptive analyses suggest that ED demand can be decomposed into many sources (such as acute bronchiolitis, asthma, or femur fractures, among many others). While some of these sources are in some sense predictable, due to following seasonal trends or being related to a measurable external factor, others are less so. For example, it is unlikely that there exists an external covariate that could help predict with precision how many people will suffer a femur fracture on any given day. Thus, there is a bound on how accurate point forecasts of ED demand can be arising from the irreducible randomness of the underlying processes that generate demand. In this work, our approach acknowledges the variability inherent to ED demand by augmenting point forecasts with prediction intervals. Empirically, we found that the 80% prediction intervals generated by the forecasting pipeline achieve near optimal performance, in that the observed number of arrivals and hospitalizations are included within the corresponding prediction interval nearly 80% of the time. This result shows that while ED demand is variable, making very precise point predictions difficult, the variability in ED demand itself is predictable, which makes it possible to generate well-performing prediction intervals.

A fundamental component of our forecasting pipeline is combining predictions from multiple base algorithms based on their empirical performance into an ensemble prediction. We found that for both point and interval forecasts, predictions based on ensembles performed better than the input predictions from each of the base algorithms. Although the performance of the three ensemble methods we investigated was comparable, Bernstein Online Aggregation was slightly more performant. In addition, the lightweight computational requirements of Bernstein Online Aggregation make it an attractive choice.

To the best of our knowledge, this is the first paper to apply conformal prediction techniques to forecasts of ED demand. We found that post-processing interval forecasts using AgACI (Zaffran et al., 2022) yielded updated intervals that achieved near optimal empirical coverage, regardless of the original algorithm used to produce the intervals. In cases where the original intervals already had good coverage, applying AgACI did not make their performance worse. Due to the simplicity and low computational burden of AgACI, this suggests that there is little downside to conformalizing intervals.

Although the final predictions from our forecasting pipeline perform well, it is notable that some of the base algorithms yield predictions that are of high quality on their own. For example, quantile regression

(A) Arrivals

	Coverage		Interval width	
Algorithm	Original	Conformal	Original	Conformal
Benchmark				
BJML	68.2%	79.6%	36.7	44.7
Base Algorithms				
ARIMA	77.8%	80.1%	35.6	38.5
ARIMAX	78.3%	80.1%	35.7	38.3
Distributional Random Forest	82.9%	80.5%	37.7	37.3
Generalized Random Forest	82.8%	80.4%	37.3	37.1
Gradient Boosted Machine	73.4%	79.6%	**29.9**	34.8
Quantile Generalized Additive Model	78.4%	80.2%	31.8	33.7
Quantile Regression	76.3%	**80.0%**	**29.9**	**33.1**
Ensemble Methods				
Bernstein Online Aggregation	**79.3%**	80.4%	32.2	33.8
Exponentially Weighted Average	77.0%	80.2%	30.8	34.0
Super Learner	76.1%	80.1%	30.1	33.5

(B) Hospitalizations

	Coverage		Interval width	
Algorithm	Original	Conformal	Original	Conformal
Benchmarks				
BJML	76.9%	**80.3%**	11.9	13.5
Base Algorithms				
ARIMAX	**80.4%**	81.0%	12.1	12.7
ARIMA	**80.4%**	81.2%	11.8	12.4
Quantile Generalized Additive Model	79.2%	81.4%	11.3	12.1
Gradient Boosted Machine	74.1%	80.5%	**10.2**	12.1
Distributional Random Forest	85.0%	80.8%	11.6	11.9
Quantile Regression	76.4%	80.7%	10.6	11.9
Generalized Random Forest	85.2%	80.7%	11.5	11.9
Ensemble Methods				
Bernstein Online Aggregation	79.3%	81.4%	10.9	**11.7**
Exponentially Weighted Average	76.4%	80.9%	10.3	11.8
Super Learner	76.4%	80.9%	**10.2**	11.8

Table 5: Performance of prediction intervals in terms of empirical coverage (with optimal level 80%) and mean interval width before and after conformalization (Step 3 of the forecasting pipeline) from each of the base algorithms and ensemble methods for (A) arrivals and (B) hospitalizations in terms of empirical coverage and mean interval width averaged over all EDs in the analysis dataset. The best-performing algorithm(s) in terms of each metric are bolded. All methods had near optimal empirical coverage after conformalization.

achieved average MAE of 9.6 and 3.4 for arrivals and hospitalizations, respectively, versus 9.3 and 3.3 for Bernstein Online Aggregation. After conformalization, quantile regression based prediction intervals had near optimal empirical coverage and interval widths comparable to the best-performing ensemble. Given these results, one may wonder why the additional step of training multiple base algorithms and applying ensembles is warranted. Indeed, depending on the context the improved performance of ensemble methods may not be worth their added computational and logistical overhead. However, in settings like ours, we note that ensemble methods allow the predictions to adapt to the possibly heterogenous trends in ED demand experienced across many EDs in a network. Indeed, we found that the best-performing base algorithm varied across EDs, and the ensemble methods generally performed as well or better than the best base algorithm. In addition, ensembles are flexible in that new prediction methods can be integrated into the system as they become available: for example, if a new forecasting method is invented, it can be added to the library of base algorithms and will only start to influence forecasts if it performs well empirically.

Our work suggests multiple directions for future research. First, going beyond the retrospective validations presented here, the forecasting pipeline could also be applied in a prospective study design to better understand its performance in real-world scenarios. Second, we focused on producing 80% prediction intervals of arrivals and hospitalizations. The forecasting pipeline could be extended to produce more than one prediction interval (for example, a collection of 80%, 90%, and 95% intervals.)

Summary points

What was already known on the topic

- Machine learning and time-series prediction for emergency departments can help improve quality of care.
- A number of related covariates, such as calendar and weather variables, may be relevant to making emergency department predictions.

What this study added to our knowledge

- Ensemble methods that combine predictions from multiple prediction algorithms yield good results for emergency department prediction tasks.
- Quantile regression methods combined with conformal inference can accurately characterize uncertainty in emergency department arrivals and hospitalizations.

Acknowledgements We would like to thank Stéphane Gaïffas, Karine Tribouley, Med Yasser Benigmim for previous work on the project and France Guyot, Nawal Derridj-Ait Younes, Reda Attia at Assistance Publique - Hôpitaux de Paris. This work was sponsored by Assistance Publique – Hôpitaux de Paris (Délégation à la Recherche Clinique et à l'Innovation).

Disclosures

Ethics approval and consent to participate This study received approval from the ethics committee (CPP Ile de France IV - IRB 00003835) under protocol number 2016/38NI.

Furthermore, regarding the individual information of participants, we obtained a waiver from the Commission Nationale Informatique et Libertés (CNIL) to do so taking into account the research methodology and in particular the difficulty of finding concerned people (authorization request no. 917254).

Consent for publication Not applicable.

Availability of data and materials The emergency department data that support the findings of this study are available from Assistance Publique - Hôpitaux de Paris but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Additional publicly available data were gathered from multiple sources: the Sentinelles network (`https://www.sentiweb.fr`) and from `data.gouv.fr`, the Iowa Environmental Mesonet (`https://mesonet.agron.iastate.edu/`), and the Global Historical Climatology Network daily (`https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily`).

Competing interests The authors declare that they have no competing interests.

Funding This research is supported by the Programme Hospitalier de Recherche Clinique – PHQ15648 (Ministère de la Santé). This research is partially supported by the Agence Nationale de la Recherche as part of the "Investissements d'avenir" program (reference ANR-19-P3IA-0001; PRAIRIE 3IA Institute).

Authors' contributions HS conducted the data analysis and drafted the manuscript. HS, AC, JJ, MW, and EB conceived the study design. PA contributed to the initiation of the study and corresponding methodology.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023. ISSN 1935-8237. doi: 10.1561/2200000101. URL http://dx.doi.org/10.1561/2200000101.

J. Scott Armstrong and Fred Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8(1):69–80, 1992. ISSN 0169-2070. doi: https://doi.org/10.1016/0169-2070(92)90008-W. URL https://www.sciencedirect.com/science/article/pii/016920709290008W.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47 (2):1148 – 1178, 2019. doi: 10.1214/18-AOS1709. URL https://doi.org/10.1214/18-AOS1709.

Antoine Augusti. Vacances scolaires par zones. Technical Report 5aeb1610c751df5402613fae, data.gouv.fr, 2023. URL https://www.data.gouv.fr/fr/datasets/vacances-scolaires-par-zones.

David Benkeser, Cheng Ju, Sam Lendle, and Mark van der Laan. Online cross-validation-based ensemble learning. *Statistics in Medicine*, 37(2):249–260, 2018. doi: https://doi.org/10.1002/sim.7320. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7320.

Kenneth Bond, Maria B. Ospina, Sandra Blitz, Marc Afilalo, Sam G. Campbell, Michael Bullard, Grant Innes, Brian Holroyd, Gil Curry, Michael Schull, and Brian H. Rowe. Frequency, determinants and impact of overcrowding in emergency departments in canada: A national survey. *Healthcare Quarterly*, 10(4): 32–40, Sep 2007. ISSN 1710-2774. URL https://www.longwoods.com/product/19312.

Kenneth P. Burnham and David Raymond Anderson. *Model selection and multimodel inference : a practical information-theoretic approach*. Springer, New York, 2nd ed edition, 2002. ISBN 0387953647; 9780387953649.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921.

Avishek Choudhury and Estefania Urena. Forecasting hourly emergency department arrival using time series analysis. *British Journal of Healthcare Management*, 26(1):34–43, 2020. doi: 10.12968/bjhc.2019.0067. URL https://doi.org/10.12968/bjhc.2019.0067.

Helene Colineaux, Fanny Pelissier, Laure Pourcel, Thierry Lang, Michelle Kelly-Irving, Olivier Azema, Sandrine Charpentier, and Sebastien Lamy. Why are people increasingly attending the emergency department? a study of the French healthcare system. *Emergency Medicine Journal*, 36(9):548–553, 2019. ISSN 1472-0205. doi: 10.1136/emermed-2018-208333. URL https://emj.bmj.com/content/36/9/548.

data.gouv.fr. Jours fériés en France. Technical Report 5b3cc551c751df4822526c1c, data.gouv.fr, 2023. URL https://www.data.gouv.fr/fr/datasets/jours-feries-en-france.

Geoffrey Ecoto, Aurélien Bibaut, and Antoine Chambaz. One-step ahead sequential super learning from short times series of many slightly dependent data, and anticipating the cost of natural disasters, 2021. URL https://arxiv.org/abs/2107.13291.

Dave R. Eitel, Scott E. Rudkin, M. Albert Malvehy, James P. Killeen, and Jesse M. Pines. Improving service quality by understanding emergency department flow: A white paper and position statement prepared for the american academy of emergency medicine. *The Journal of Emergency Medicine*, 38(1):70–79, 2010.

Matteo Fasiolo, Simon N. Wood, Margaux Zaffran, Raphaël Nedellec, and Yannig Goude. Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, 116(535):1402–1412, 2020. doi: 10.1080/01621459.2020.1725521. URL https://doi.org/10.1080/01621459.2020.1725521.

Matteo Fasiolo, Simon N. Wood, Margaux Zaffran, ël Nedellec, and Yannig Goude. qgam: Bayesian non-parametric quantile regression modeling in R. *Journal of Statistical Software*, 100(9):1–31, 2021a. doi: 10.18637/jss.v100.i09.

Matteo Fasiolo, Simon N. Wood, Margaux Zaffran, Raphaël Nedellec, and Yannig Goude. Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, 116(535):1402–1412, 2021b. doi: 10.1080/01621459.2020.1725521. URL https://doi.org/10.1080/01621459.2020.1725521.

Antoine Flahault, Thierry Blanchon, Yves Dorléans, Laurent Toubiana, Jean-François Vibert, and Alain-Jacques Valleron. Virtual surveillance of communicable diseases: a 20-year experience in France. *Statistical Methods in Medical Research*, 15(5):413–421, 2006. doi: 10.1177/0962280206071639. URL https://doi.org/10.1177/0962280206071639. PMID: 17089946.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001. doi: 10.1214/aos/1013203451. URL https://doi.org/10.1214/aos/1013203451.

Jessica Gacki-Smith, Altair M. Juarez, Lara Boyett, Cathy Homeyer, Linda Robinson, and Susan L. MacLean. Violence against nurses working in us emergency departments. *JONA: The Journal of Nursing Administration*, 39(7/8), 2009. ISSN 0002-0443. URL https://journals.lww.com/jonajournal/fulltext/2009/07000/violence_against_nurses_working_in_us_emergency.9.aspx.

Pierre Gaillard, Yannig Goude, Laurent Plagne, Thibaut Dubois, and Benoit Thieurmel. *opera: Online Prediction by Expert Aggregation*, 2023. URL http://pierre.gaillard.me/opera.html. R package version 1.2.1.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0d441de75945e5acbc865406fc9a2559-Paper.pdf.

Isaac Gibbs and Emmanuel Candès. Conformal inference for online prediction with arbitrary distribution shifts, 2022.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011a. ISSN 01621459. URL http://www.jstor.org/stable/41416407.

Tilmann Gneiting. Quantiles as optimal point forecasts. *International Journal of Forecasting*, 27(2):197–207, 2011b. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2009.12.015. URL https://www.sciencedirect.com/science/article/pii/S0169207010000063.

Jean-Yves Grall. rapport sur la territorialisation des activites d'urgences [report on the territorialization of emergency activities]. Technical report, Paris: Ministère des Affaires sociales de la Santé et des Droits des femmes., 2015.

Brandon Greenwell, Bradley Boehmke, Jay Cunningham, and GBM Developers. *gbm: Generalized Boosted Regression Models*, 2022. URL https://CRAN.R-project.org/package=gbm. R package version 2.1.8.1.

Muhammet Gul and Erkan Celik. An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. *Health Systems*, 9(4):263–284, 2020. doi: 10.1080/20476965.2018.1547348. URL https://doi.org/10.1080/20476965.2018.1547348.

Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297 – 310, 1986. doi: 10.1214/ss/1177013604. URL https://doi.org/10.1214/ss/1177013604.

I Higginson, J Whyatt, and K Silvester. Demand and capacity planning in the emergency department: how to do it. *Emergency Medicine Journal*, 28(2):128–135, 2011. ISSN 1472-0205. doi: 10.1136/emj.2009.087411. URL https://emj.bmj.com/content/28/2/128.

Nathan R. Hoot and Dominik Aronsky. Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine*, 52(2):126–136.e1, 2008. ISSN 0196-0644. doi: https://doi.org/10.1016/j.annemergmed.2008.03.014. URL https://www.sciencedirect.com/science/article/pii/S0196064408006069.

Rob Hyndman, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. *forecast: Forecasting functions for time series and linear models*, 2023. URL https://pkg.robjhyndman.com/forecast/. R package version 8.21.

Rob J. Hyndman and Yeasmin Khandakar. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3):1–22, 2008. doi: 10.18637/jss.v027.i03. URL https://www.jstatsoft.org/index.php/jss/article/view/v027i03.

Shancheng Jiang, Qize Liu, and Beichen Ding. A systematic review of the modelling of patient arrivals in emergency departments. *Quantitative Imaging in Medicine and Surgery*, 13(3), 2022. ISSN 2223-4306. URL https://qims.amegroups.com/article/view/102522.

Roger Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005. doi: 10.1017/CBO9780511754098.

Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng, editors. *Handbook of Quantile Regression*. Chapman and Hall/CRC, 1 edition, 2017. doi: 10.1201/9781315120256.

Shan W. Liu, Yuchiao Chang, Joel S. Weissman, Richard T. Griffey, James Thomas, Suvd Nergui, Azita G. Hamedani, Carlos A. Camargo Jr., and Sara Singer. An empirical assessment of boarding and quality of care: Delays in care among chest pain, pneumonia, and cellulitis patients. *Academic Emergency Medicine*, 18(12):1339–1348, 2011. doi: https://doi.org/10.1111/j.1553-2712.2011.01082.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1553-2712.2011.01082.x.

Paula Maddigan and Teo Susnjak. Forecasting patient demand at urgent care clinics using machine learning, 2022. URL https://arxiv.org/abs/2205.13067.

Spyros Makridakis. Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 9(4):527–529, 1993. ISSN 0169-2070. doi: https://doi.org/10.1016/0169-2070(93)90079-3. URL https://www.sciencedirect.com/science/article/pii/0169207093900793.

Izabel Marcilio, Shakoor Hajat, and Nelson Gouveia. Forecasting daily emergency department visits using calendar variables and ambient temperature readings. *Academic Emergency Medicine*, 20(8):769–777, 2013. doi: https://doi.org/10.1111/acem.12182. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/acem.12182.

Jordi McKenzie. Mean absolute percentage error and bias in economic forecasting. *Economics Letters*, 113(3):259–262, 2011. ISSN 0165-1765. doi: https://doi.org/10.1016/j.econlet.2011.08.010. URL https://www.sciencedirect.com/science/article/pii/S0165176511003119.

Matthew J. Menne, Imke Durre, Russell S. Vose, Byron E. Gleason, and Tamara G. Houston. An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29(7):897 – 910, 2012. doi: https://doi.org/10.1175/JTECH-D-11-00103.1. URL https://journals.ametsoc.org/view/journals/atot/29/7/jtech-d-11-00103_1.xml.

Claire Morley, Maria Unwin, Gregory M. Peterson, Jim Stankovich, and Leigh Kinsman. Emergency department crowding: A systematic review of causes, consequences and solutions. *PLOS ONE*, 13(8):1–42, 08 2018. doi: 10.1371/journal.pone.0203316. URL https://doi.org/10.1371/journal.pone.0203316.

John C. Moskop, David P. Sklar, Joel M. Geiderman, Raquel M. Schears, and Kelly J. Bookman. Emergency department crowding, part 1—concept, causes, and moral consequences. *Annals of Emergency Medicine*, 53(5):605–611, 2009.

Jesse M. Pines, Joshua A. Hilton, Ellen J. Weber, Annechien J. Alkemade, Hasan Al Shabanah, Philip D. Anderson, Michael Bernhard, Alessio Bertini, André Gries, Santiago Ferrandiz, Vijaya Arun Kumar, Veli-Pekka Harjola, Barbara Hogan, Bo Madsen, Suzanne Mason, Gunnar Öhlén, Timothy Rainer, Niels Rathlev, Eric Revue, Drew Richardson, Mehdi Sattarian, and Michael J. Schull. International perspectives on emergency department crowding. *Academic Emergency Medicine*, 18(12):1358–1370, 2011. doi: https://doi.org/10.1111/j.1553-2712.2011.01235.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1553-2712.2011.01235.x.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

Bahman Rostami-Tabar, Jethro Browell, and Ivan Svetunkov. Probabilistic forecasting of hourly emergency department arrivals. *Health Systems*, 0(0):1–17, 2023. doi: 10.1080/20476965.2023.2200526. URL https://doi.org/10.1080/20476965.2023.2200526.

Melanie Roussel, Dorian Teissandier, Youri Yordanov, Frederic Balen, Marc Noizet, Karim Tazarourte, Ben Bloom, Pierre Catoire, Laurence Berard, Marine Cachanado, Tabassome Simon, Said Laribi, Yonathan Freund, and FHU IMPEC–IRU SFMU Collaborators. Overnight Stay in the Emergency Department and Mortality in Older Patients. *JAMA Internal Medicine*, 11 2023. ISSN 2168-6106. doi: 10.1001/jamainternmed.2023.5961. URL https://doi.org/10.1001/jamainternmed.2023.5961.

Maëlle Salmon. *riem: Accesses Weather Data from the Iowa Environment Mesonet*, 2023. https://docs.ropensci.org/riem/, https://github.com/ropensci/riem.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421, jun 2008. ISSN 1532-4435.

Herbert Susmann, Antoine Chambaz, and Julie Josse. Adaptiveconformal: An r package for adaptive conformal inference, 2023.

Alain-Jacques Valleron, Elisabeth Bouvet, Philippe Garnerin, Juan Ménarès, Isabelle Heard, Sylvia Letrait, and Jacques Lefaucheux. A computer network for the surveillance of communicable diseases: the French experiment. *American Journal of Public Health*, 76(11):1289–1292, 1986. doi: 10.2105/AJPH.76.11.1289. URL https://doi.org/10.2105/AJPH.76.11.1289. PMID: 3766824.

Mark J. van der Laan, Eric C Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007. doi: doi:10.2202/1544-6115.1309. URL https://doi.org/10.2202/1544-6115.1309.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

M Wargon, B Guidet, T D Hoang, and G Hejblum. A systematic review of models for forecasting the number of emergency department visits. *Emergency Medicine Journal*, 26(6):395–399, 2009. ISSN 1472-0205. doi: 10.1136/emj.2008.062380. URL https://emj.bmj.com/content/26/6/395.

Olivier Wintenberger. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106(1):119–141, Jan 2017. ISSN 1573-0565. doi: 10.1007/s10994-016-5592-6. URL https://doi.org/10.1007/s10994-016-5592-6.

Margaux Zaffran, Olivier Feron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25834–25866. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zaffran22a.html.

Xinxing Zhao, Joel Weijia Lai, Andrew Fu Wah Ho, Nan Liu, Marcus Eng Hock Ong, and Kang Hao Cheong. Predicting hospital emergency department visits with deep learning approaches. *Biocybernetics and Biomedical Engineering*, 42(3):1051–1065, 2022. ISSN 0208-5216. doi: https://doi.org/10.1016/j.bbe.2022.07.008. URL https://www.sciencedirect.com/science/article/pii/S0208521622000729.

Alexander Zlotnik, Ascensión Gallardo-antolín, Miguel Cuchí Alfaro, María Carmen Pérez Pérez, and Juan Manuel Montero Martínez. Emergency department visit forecasting and dynamic nursing staff allocation using machine learning techniques with readily available open-source software. *CIN: Computers, Informatics, Nursing*, 33(8), 2015. ISSN 1538-2931. URL https://journals.lww.com/cinjournal/Fulltext/2015/08000/Emergency_Department_Visit_Forecasting_and_Dynamic.7.aspx.

Domagoj Ćevid, Loris Michel, Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022. URL http://jmlr.org/papers/v23/21-0585.html.

7 Appendix

7.1 Covariates

Let y_t be the outcome (arrivals or hospitalizations) in an ED on day t. The following covariates are used for predicting y_t (one-day lag refers to day $t-1$):

- y_{t-1}: one-day lagged outcome.
- y_{t-7}: one-day lagged outcome.
- $y_{t-1} - y_{t-2}$: difference in previous two outcomes.
- Indicator of national holiday. National holidays are January 1, Easter Monday, May 1st, May 8th, Ascension Day, Whit Monday, July 14th, Assumption Day, All Saints Day, November 11th, and December 25th.
- Indicator of day following a national holiday.
- Indicator of a Friday following a national holiday.
- Separate indicators for January 1st, July 14th, December 25th, December 31st.
- Indicator of school vacation.
- One-week lagged incidence in Île-de-France of flu-like maladies, diarrhea, and chickenpox.
- One-day lagged values of maximum observed temperature, minimum observed temperature, and observed precipitation.

7.2 Overview of base algorithms

Time-series models The Autoregressive integrated moving average (ARIMA) model is widely used for time-series forecasting. An ARIMA model is defined by the number of lags p, order of differencing d, and order of moving differences q, where each parameter is a positive integer. Formally, an ARIMA(p, d, q) process is characterized by

$$\phi(B)(1 - B^d)y_t = c + \theta(B)\varepsilon_t,$$

where B is the backshift operator (that is, $B^a y_t = y_{t-a}$), ϕ and θ are polynomials of order p and q, and (ε_t) is an independent and identically distributed (iid) white noise process with mean 0 and variance σ^2. The parameter values were chosen in a model selection procedure based on unit-root tests and minimizing the corrected AIC of the considered models (Burnham and Anderson, 2002; Hyndman and Khandakar, 2008). Prediction intervals were defined using the estimated variance of the ARIMA process in the form $[\hat{y}_t - q_{0.1}\hat{\sigma}, \hat{y}_t + q_{0.9}\hat{\sigma}]$, where $\hat{y}_t$ is the ARIMA point estimate, q_α is the α-quantile of the standard normal distribution, and $\hat{\sigma}$ is the estimated ARIMA variance. We also used ARIMA with covariates (referred to as ARIMAX) as a candidate algorithm, with holidays, school vacation, weather, and public health surveillance covariates as described previously (Section 7.1).

Regression approach Following a typical regression setup, quantile regression assumes a linear relationship between a vector of covariates $\boldsymbol{x}_t$ and the outcome y_t. Rather than minimizing the mean squared error of the predictions, as in standard linear regression, in quantile regression the mean quantile loss is minimized. All variables described in Section 7.1 were used as covariates. In addition, one-day and one-week lagged values of the outcome were included as additional covariates to capture temporal autocorrelation, as well as the difference in the prior two days outcomes (designed to capture whether the outcome is increasing or decreasing). Formally, we assume the linear relationship

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-7} + \alpha_3(y_{t-1} - y_{t-2}) + \boldsymbol{\beta}^\top \boldsymbol{X}_t + \epsilon_t$$

where y_t is the outcome (arrivals or hospitalizations) on day t, $\boldsymbol{X}_t$ is a vector of covariates with associated coefficients $\boldsymbol{\beta}$, and ϵ_t is a residual.

A potential shortcoming of quantile regression is that it assumes a linear relationship between covariates and outcome. However, it is not clear a-priori that this linearity assumption is warranted in our setting. For example, we might expect maximum daily temperature to have a non-linear relationship with emergency department arrivals. Generalized Additive Models (GAMs) allow for non-linear relationships between covariates and outcome as modeled by smooth functions, such as splines (Hastie and Tibshirani, 1986). The GAM framework has been extended to quantile estimation, with fast computational procedures available (Fasiolo et al., 2021b,a). For our application, we used the same covariates as for quantile regression. Spline-based smoothers were used to model the relationship between the continuous covariates and the outcome.

Machine learning approach Several machine learning approaches were also included in the library of base algorithms, focusing particularly on flexible tree-based approaches. Generalized Random Forest (GRF) framework is one such method method for growing forests of decision trees that can be used for estimating diverse statistical functionals (Athey et al., 2019). We use a version of GRF, Quantile Random Forests (QRF), tailored to estimating conditional quantiles. Distributional Random Forests (DRF) are a further generalization of the Random Forest algorithm that yields estimates of the full conditional distribution of an outcome from which any statistical summary measure, such as quantiles, can be calculated (Ćevid et al., 2022). As a final machine learning approach we considered Gradient boosting, a general technique for combining many weak learners (such as decision trees) into an ensemble estimator (Friedman, 2001). The underlying idea is to iteratively improve an estimator by training a new learner on the errors of the current model. We apply gradient boosting with the quantile loss function, thus estimating conditional quantiles of the forecast distribution.

7.3 Base algorithm implementation details

Base algorithm details:

- **ARIMA and ARIMAX:** We used the `auto.arima` model selection procedure for ARIMA, which chooses the ARIMA specification based on unit root tests and corrected AIC (see Burnham and Anderson (2002); the model selection algorithm is described in Hyndman and Khandakar (2008)). No seasonality was allowed in the model selection procedure. For ARIMAX, all covariates except the lagged outcomes (y_{t-1}, y_{t-7}) and differenced outcomes ($y_{t-1} - y_{t-2}$)) were included.
- **Quantile Regression:** default parameters from the `quantreg` package were used. All covariates were included.
- **Quantile Generalized Additive Models:** default parameters from the `qgam` package were used. All covariates were included. Spline smoothers were applied to all variables that were not indicators.
- **Generalized Random Forest:** default tuning parameters from the `grf` package were used. All covariates were included.
- **Directional Random Forest:** default tuning parameters from the `drf` package were used. All covariates were included.
- **Gradient Boosted Machine:** default tuning parameters from the `gbm` package were used. All covariates were included.

7.4 Additional results

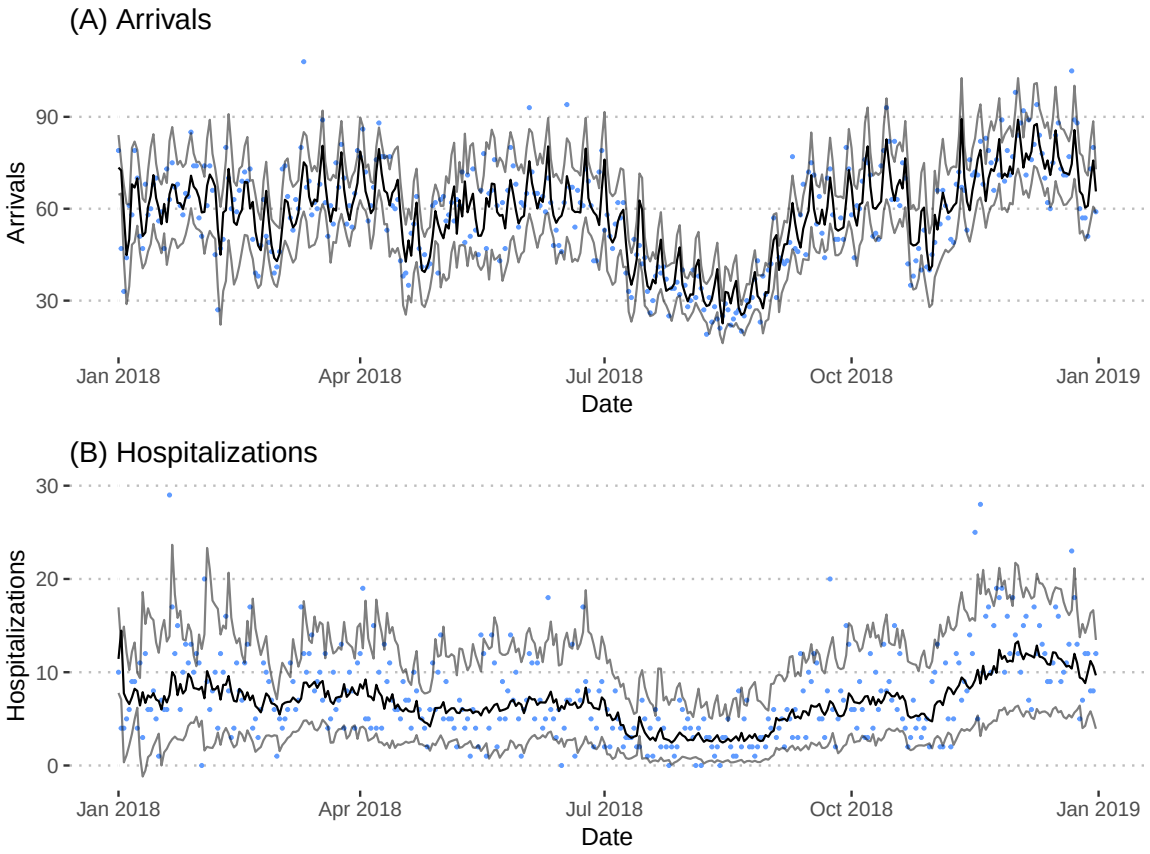

Figure 6: One-day ahead point predictions (black line) and 80% prediction intervals (gray lines) based on Bernstein Online Aggregation for arrivals (A) and hospitalizations (B) for one ED in the analysis dataset. Observed number of arrivals and hospitalizations are shown as blue points.

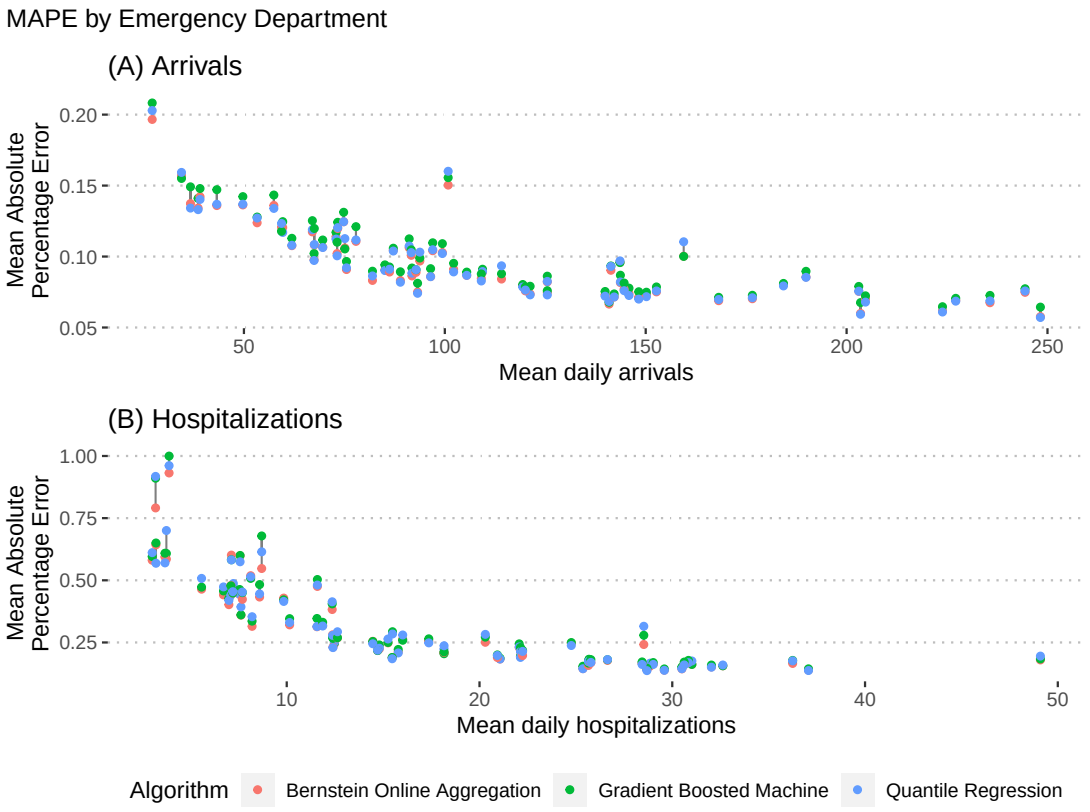

Figure 7: Mean absolute percentage error (MAPE) of point predictions for (A) arrivals and (B) hospitalizations from Bernstein Online Aggregation, Gradient Boosted Machine, and Quantile Regression. Each point shows the performance of an algorithm in one ED, and points from the same ED are connected by a line. The x-axis is the mean daily arrivals and hospitalizations across all days in the analysis dataset.

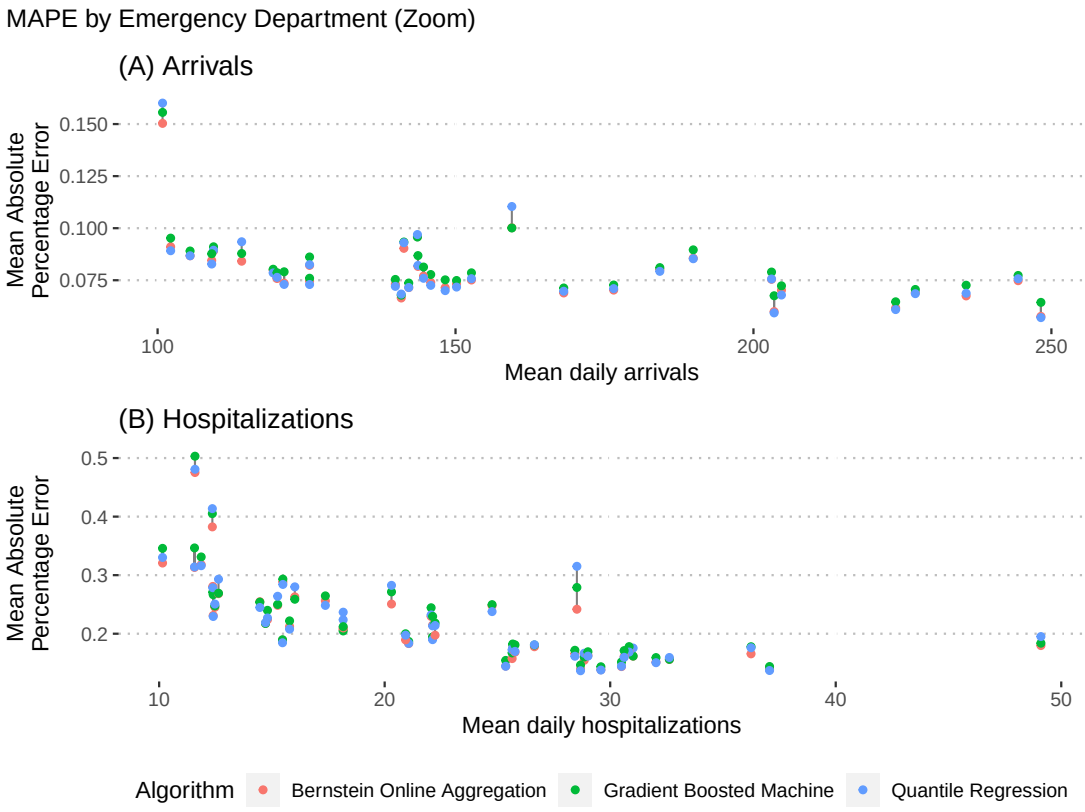

Figure 8: Mean absolute percentage error (MAPE) of point predictions for (A) arrivals and (B) hospitalizations from Bernstein Online Aggregation, Gradient Boosted Machine, and Quantile Regression. In (A) only EDs with mean daily arrivals greater than 100 are shown, and in (B) only EDs with mean daily hospitalizations greater than 10 are shown. Each point shows the performance of an algorithm in one ED, and points from the same ED are connected by a line. The x-axis is the mean daily arrivals and hospitalizations across all days in the analysis dataset.