A unified machine learning approach to time series forecasting applied to demand at emergency departments

Michaela A.C. Vollmer[1]* ![ORCID], Ben Glampson[2], Thomas Mellan[1], Swapnil Mishra[1], Luca Mercuri[2], Ceire Costello[1], Robert Klaber[2], Graham Cooke[2], Seth Flaxman[3] and Samir Bhatt[1,4]

Abstract

Background: There were 25.6 million attendances at Emergency Departments (EDs) in England in 2019 corresponding to an increase of 12 million attendances over the past ten years. The steadily rising demand at EDs creates a constant challenge to provide adequate quality of care while maintaining standards and productivity. Managing hospital demand effectively requires an adequate knowledge of the future rate of admission. We develop a novel predictive framework to understand the temporal dynamics of hospital demand.

Methods: We compare and combine state-of-the-art forecasting methods to predict hospital demand 1, 3 or 7 days into the future. In particular, our analysis compares machine learning algorithms to more traditional linear models as measured in a mean absolute error (MAE) and we consider two different hyperparameter tuning methods, enabling a faster deployment of our models without compromising performance. We believe our framework can readily be used to forecast a wide range of policy relevant indicators.

Results: We find that linear models often outperform machine learning methods and that the quality of our predictions for any of the forecasting horizons of 1, 3 or 7 days are comparable as measured in MAE. Our approach is able to predict attendances at these emergency departments one day in advance up to a mean absolute error of ± 14 and ± 10 patients corresponding to a mean absolute percentage error of 6.8% and 8.6% respectively.

Conclusions: Simple linear methods like generalized linear models are often better or at least as good as ensemble learning methods like the gradient boosting or random forest algorithm. However, though sophisticated machine learning methods are not necessarily better than linear models, they improve the diversity of model predictions so that stacked predictions can be more robust than any single model including the best performing one.

Keywords: Machine learning, Time series analysis, Emergency department demand, Ensemble predictions

*Correspondence: m.vollmer@imperial.ac.uk
Seth Flaxman is a joint senior author.
[1]MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, UK
Full list of author information is available at the end of the article

Background

In 2019 there were 25.6 million attendances at emergency departments (EDs) in the UK, corresponding to 70,231 patients attending every day [1]. The National Health Service (NHS) trusts across England are under very high pressure to maintain current standards and quality of care [2]. In fact, the rate of attendance has grown by 4.8% since 2018 and by 10.6% over the last 5 years meaning that it is increasing at a rate faster than population growth thus putting high pressure on our health care system. Failure to make provisions for surges in demand can lead to overcrowding, which in turn has been linked to multiple adverse patient outcomes such as unfavourable patient satisfaction, poor quality of care and diseconomies of scale [3]. In order to decrease overcrowding, the NHS introduced a new operational standard in 2010, commonly-known as the "four hour target", requiring that at least 95% of patients attending EDs should either be discharged, admitted or transferred within 4 hours of arrival. However, this target has not been met since 2014 and failure rates have reached a new high in January 2020 with 18% of patients at EDs waiting longer than four hours despite the fact that the overall number of attendances was lower in January 2020 than in January 2019. In fact 15.3% of patients spent more than 4 hours in EDs in 2019 compared to 11.9% last year and 5.5% five years ago, making 2019 the year with the worst annual performance on record [1]. A shortage of staff is not the predominant cause for long waiting times and low quality of care [4], rather, it is that the correct *type* of staffing is not matched with patient demand. This inefficiency in staffing can have substantial impacts such as the 86,264 of elective surgeries that had to be cancelled for non-clinical reasons on the day the patient was due to arrive in 2019. These cancellations leave NHS hospital trusts with lost costs of surgeons, anaesthetists and nurses as well as surgical session time and theatre capacity. Moreover, in the same year, the percentage of patients who had not been treated within 28 days of cancellation decreased from 9.8% in 2018 to 8.8% in 2019, however, still failing one of the NHS's improvement objectives. In addition the NHS England advises that a 85% bed occupancy rate is the maximum safe level of occupancy and it advises that trusts should try and keep bed occupancy below 92%. However, 126 out of 170 trusts recorded bed occupancy above 85%, 58 trusts rates above 92% in the third quarter of 2019, eight trusts had occupancy above 98% and one trust recorded 100% bed occupancy (NHS SitRep).

Overall the NHS in England has spent around £130 billion in 2018/19 on the delivery of health services [5]. A major fraction, 44.9% in the financial year 2016-2017 [6], of this spending is due to the 1.2 million people employed by NHS hospital and community health services. Between February 2018 and 2019 the number of doctors alone rose by 2.5% and by 10% during the past five years [7]. This emphasises the fact that long waiting times may not simply result from a shortage of staff. Nonetheless, every year high agency and staffing costs are required to cover for staffing shortages.

Given the importance of health care optimisation, procedures to help the hospitals balance the right staffing numbers deployed in the most effective way is paramount. This not only has implications for patient outcomes but for general efficiency within the NHS. This paper provides a means to improve health care delivery through a predictive understanding of the rates of admission and the duration of stay at EDs (collectively referred to as "demand"). In particular we provide statistical techniques that can help facilitate an operationally relevant prediction of demand for use in day to day planning. We also contribute to a better understanding of the reasons for demand and highlight why this understanding is fundamental to the delivery of high quality care.

Previous work

Despite its importance, methods for predicting rates of admission and understanding underlying dynamics have not been studied extensively in the literature. Existing methods have been limited to the application of classical time series forecasting methods. In an Australian study, Boyle et al. [8] predict monthly, daily and four hourly demand at 27 EDs in Queensland. Taking public holidays into account as predictors, the authors used an autoregressive integrated moving average (ARIMA) model as well as regression and exponential smoothing methods to predict demand up to a mean absolute percentage error (MAPE) of 11% for daily admissions. McCarthy et al. [9] predicted hourly presentations to American emergency departments while including several temporal, weather and patient factors on the number of hourly arrivals in order to characterise behaviour at EDs. Jones et al. [10] used seasonal autoregressive integrated moving average (SARIMA), time series regression, exponential smoothing and artificial neural network models to forecast demand. The authors had access to two years of data and made predictions ranging from 1 to 30 days in advance for three hospitals in the USA. The authors found seasonal and weekly patterns to be most important for an accurate prediction and obtained a Mean Absolute Predictive Error (MAPE) of $9 - 13$% depending on the facility. Champion et al. [11] performed an analysis of monthly demand at an emergency department in regional Victoria. They applied exponential smoothing and Box-Jenkins methods to five years worth of admissions data. Hoot et al. [12] carried out a discrete event simulation in order to obtain forecasts for 2, 4, 6 and 8 hours into the future. The authors analysed waiting times, length of stay and bed occupancy.

Our contribution

We develop a novel, predictive framework to understand the temporal dynamics of hospital demand and we apply an exhaustive statistical analysis to daily presentations at EDs at St Mary's and Charing Cross Hospitals, evaluating a range of standard time series and machine learning approaches and ultimately developing our own unique approach. In contrast to existing studies, we do not only focus on the application of time series algorithms in order to characterise demand but develop a generic procedure that allows us to compare and combine both time series and machine learning algorithms in order to obtain an informative, more appropriate and consistently accurate approach to the prediction of demand. Our models have the ability to be retrained regularly and efficiently and are therefore a powerful tool for online platforms and near real time prediction. Using novel data from electronic logging systems from eight years of daily presentations to EDs at St Mary's and Charing Cross Hospitals in London, we construct a model that predicts the number of daily arrivals to both hospitals. Our analysis accounts for seasonal fluctuations, daily observed weather data and specific, pre-planned events indicated by staff at both EDs such as the yearly Notting Hill Carnival. Using our procedure can help the hospitals with the provision of the right staffing numbers and deploying resources in the most effective way.

Data sources

The St Mary's and Charing Cross Hospitals are part of the Imperial College Healthcare NHS Trust, one of 228 NHS hospital trusts in England. St Mary's Hospital is the major acute care hospital for North West London housing a major trauma centre. Its ED has faced an average demand of 208 patients (with a maximum of 289 and a minimum of 99) every day since 2011. Charing Cross Hospital includes the serious injuries centre for West London as well as a hyper acute stroke unit. On average there have been 106 (with a maximum of 177 and a minimum of 60) daily attendances at the ED since 2011. For our analysis, we had access to electronic data records of the number of daily attendances at the EDs for both hospitals from 2011 to 2018 (see Fig. 1). In order to investigate the demand dynamics, we also collected data on school [13] and bank holidays [14], as well as on the weather and Google search volume for the word "flu" [15, 16] (see the Appendix for more details). Finally, experienced staff at the EDs of both hospitals provided us with a list of specific *known* events in the locality that cause surge in demand (e.g. the Notting Hill carnival - an annual festival taking place in the catchment area of the hospital).
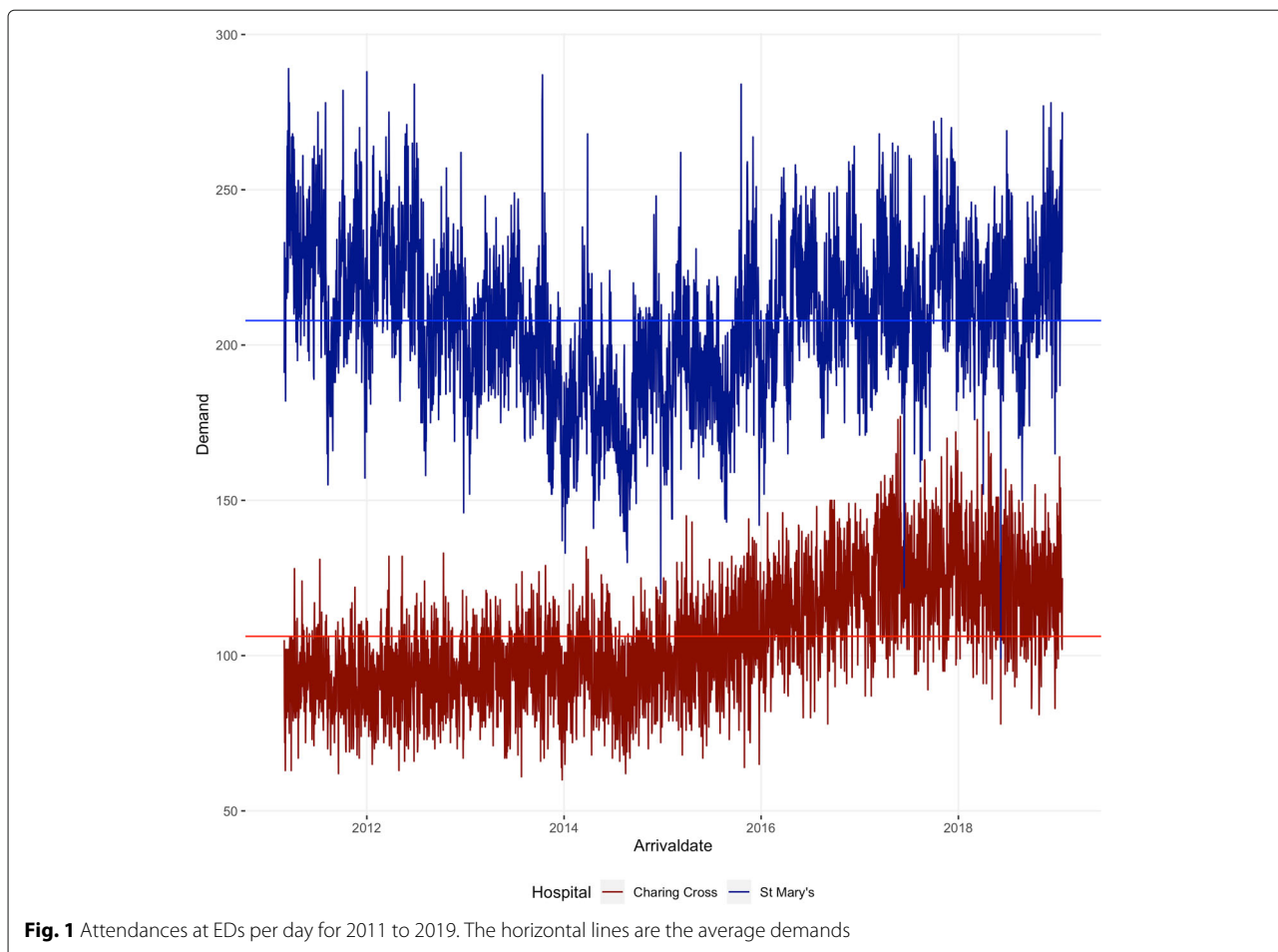
Exploratory data analysis

A central aim of this paper is not only to predict hospital demand accurately but also attempt to understand the factors driving hospital demand. Daily data is driven by a complex web of exogenous variables, many of which are related to seasonal patterns or trends. Both time series depicted in Fig. 1 show a strong underlying trend. In the case of Charing Cross Hospital this trend is clearly upwards while it goes downwards first for St Mary's Hospital due to a change of the hospital's infrastructure, see Fig. 2. There is also clear seasonality, the monthly attendance at St Mary's Hospital shows a very clear monthly periodic pattern with troughs in January, April and August and a rise in attendance during the winter months (likely due to increases in acute respiratory infections), see Fig. 3. Figure 3 also shows indications that bank or school holidays have a strong influence on the number of ED admissions together with the flu season. It also shows that the flu seasons contribution to increased demand runs well into spring. While the monthly attendance at Charing Cross Hospital also shows some periodic behaviour, it is not as strong, see Fig. 4. It is therefore useful to note that dynamics differ even from geographically close hospitals with overlapping catchments. Both series show clear day-of-week patterns, characterised by a strong autocorrelation with respect to their lagged values of order 7, see Fig. 5. Mondays have the highest volume of attendances at both hospitals while attendance reaches its minimum during weekends. This finding validates and confirms other studies on hospital demand [17, 18].

Methods

We focus on forecasting demand one, three, and seven days into the future. These particular forecasting intervals are relevant as they allow the hospitals to take action by using short term measures such as the cancellation of elective surgeries or the hiring of additional staff through agencies.

We use two different kinds of algorithms for our predictions: traditional time series and machine learning algorithms [19]. A discrete time series is a sequence of data points in chronological order divided into regular time intervals. The fundamental assumption behind both algorithms is that data points that are close to each other in time show a similar behaviour and that there is a dependency between data points at the same position of the time interval, e.g. same time of the year or same day of the week. Time series algorithms use both the chronology of the events and the specified interval in order to make inference and split the time series into different linear components such as seasonality, trend and a residual. The residual, which is assumed to contain some correlative structure, is usually modelled using an autoregressive stochastic process or exponential decay where future values are predicted based on past values [19]. In contrast, machine learning algorithms specify a broad function

Fig. 1 Attendances at EDs per day for 2011 to 2019. The horizontal lines are the average demands
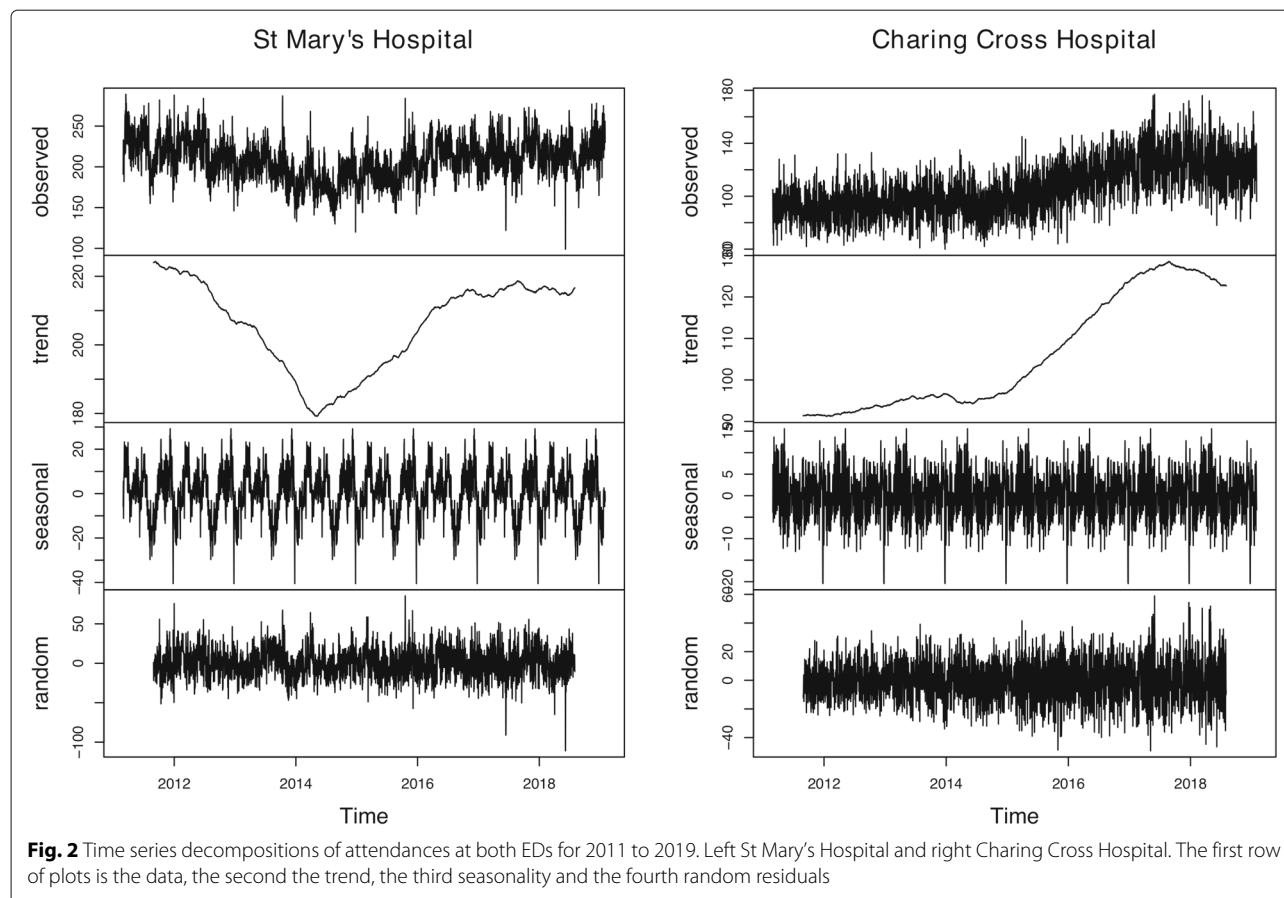
class (such as trees or smooth curves) with sufficient capacity to learn complex functions. These algorithms learn from data balancing function complexity with predictive accuracy. For both sets of algorithms we create a model (design) matrix containing explanatory variables. For the time series algorithms only lagged demand from previous time points were used. For the machine learning algorithm lagged demand values were used alongside other covariates (see Table 1). As predictors we have chosen past values such as demand on the previous day, last week and the average of the past week as well as indicators for bank holidays and school holidays. Moreover, we use data from some of the surrounding weather stations on precipitation, minimal and maximal temperature as covariates. Finally, we use search engine query data as a covariate, as it has proven to be a very efficient measure for the detection of influenza epidemics [20]. Table 1 shows a few rows of our model matrix.

Below, we summarise the time series and machine learning algorithms that we have considered (for detailed mathematical information we refer the reader to [21]:

Time series algorithms

1 ARIMA - AutoRegressive Integrated Moving Average [21] An ARIMA model consists of three parts: The autoregressive component (AR) referring to the fact that the indicator of interest is regressed on its own previous values (i.e. current values of demand depend on past values of demand), the integrated (I) part representing one or several differencing steps to make the time series stationary and the moving average (MA) component indicating that the regression error is a linear combination of past error terms.

2 ETS - Exponential smoothing methods [21] In general exponential smoothing refers to forecasting methods which also regress on lagged values of the target variable. However, it uses exponentially decreasing weights for past observations. The ETS model we employ uses exponential smoothing for error, trend and seasonality.

3 STLM - a seasonal decomposition of time series by LOESS (STL)a seasonal decomposition of time series by LOESS (STL) STLM is another type of

Fig. 2 Time series decompositions of attendances at both EDs for 2011 to 2019. Left St Mary's Hospital and right Charing Cross Hospital. The first row of plots is the data, the second the trend, the third seasonality and the fourth random residuals

exponential smoothing model. The time series is decomposed into its seasonal components using LOESS (locally estimated scatterplot smoothing) before exponential smoothing is used to model the error and trend component of the time series. Finally the series is re-seasonalized.
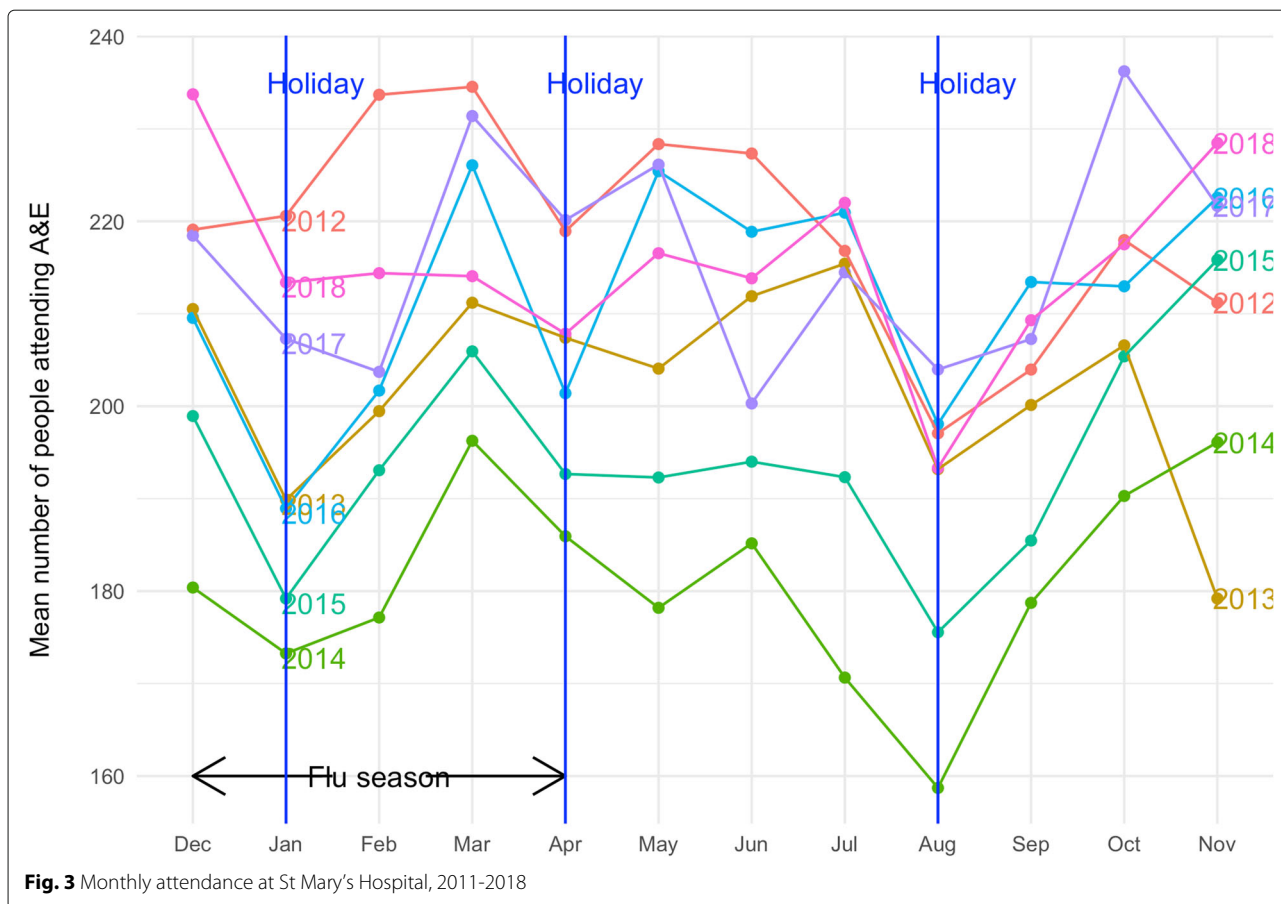
4 StructTS - Structural Time Series Model [21] A StructTS model is formulated directly in terms of unobserved components, such as trends, seasonality and exogenous factors that have a natural interpretation. The StructTS model forms a state space which makes it similar to the ARIMA and ETS models.

Machine learning algorithms

1 glmnet [22] is a penalised generalized linear model with built-in variable selection. The glmnet model is an extension of the generalised linear model in which bias (penalty/weight decay/regulariser) is introduced in the form of a mixture penalty consisting of the parameter ℓ_1 and ℓ_2 norms. The magnitude of this penalty is tuned to balance overfitting vs.

underfitting, with the goal of reducing variance by introducing some bias.

2 ranger [23] is a fast implementation of random forests [24]. Random forests are a bagged ensemble of decision trees. Random forests use bootstrapped random subsets of the covariates variables to build decision trees based on these subsets. Random forests prevent overfitting by averaging to reduce variance.

3 Gradient Boosting Machines (gbm) [25] are a generalized boosted regression model. In contrast to random forests which builds a collection of multiple independent decision trees, the decision trees created by gradient boosting machines depend additively on each other. Each new tree is added to an ensemble by improving the previous trees residual error (the functional negative gradient). Hyperparameters are tuned to balance over and under fitting.

4 k-nearest neighbours (k-NN) [26] is a classic supervised learning algorithm based on the idea that data points that are close to each other in the space of covariates should have similar predictions. Hence, to make a prediction at a given location in covariate

Fig. 3 Monthly attendance at St Mary's Hospital, 2011-2018

space, an average of the labels of the k nearest neighbours is taken. The disadvantage of this algorithm is that it slows down quickly once the volume of data points increases.

Stacked regression
All the above algorithms have strengths and weaknesses. It is therefore challenging to choose a single model for prediction. We create a consensus model by adopting stacked regression, a particularly effective ensemble approach. The idea of stacked regression is to combine the diversity and strengths of multiple algorithms into a single model with a better overall ability to generalise. We choose a linear stacking model subject to convex combination constraints [27]. Following [27, 28] we train a linear stacker on cross validation predictions of the individual time series and machine learning models. This procedure helps us avoid selecting models that overfit to the data. Stacking models is not only empirically motivated, but has a strong theoretical backing and has been proven to perform asymptotically exactly as well (by some loss metric) as the best possible choice for the given data set among the family of weighted combinations of the estimators. Stacking has also been showing to work in a variety of settings [29].

Validation and evaluation
Developing an algorithm with the best possible forecast accuracy is the main goal of this study. However, care must be taken to ensure that these forecasts are not simply overfitting to noise in the data but are accurate and can truly forecast to unseen data. A key innovation of this paper is the development of a novel general purpose time series cross validation procedure to ensure that:

a all algorithms are evaluated fairly and equally,
b the same data is used in all algorithms, and
c forecast errors are completely blind to the held-out data (i.e. exactly as if the model was being used in a real forecasting setting).

Temporal or time series cross-validation [21] is a method to split the data into testing and training sets in order to account for temporal structure in the data. The main idea is that each test set only consists of a forecasting window of one day which lies one, three or seven days in the future while the corresponding training set consists of a number of observations prior to the forecasting window. The origin can either be fixed so that the length of the training window grows by one, three, or seven days for each new test set or it can move forward so that the
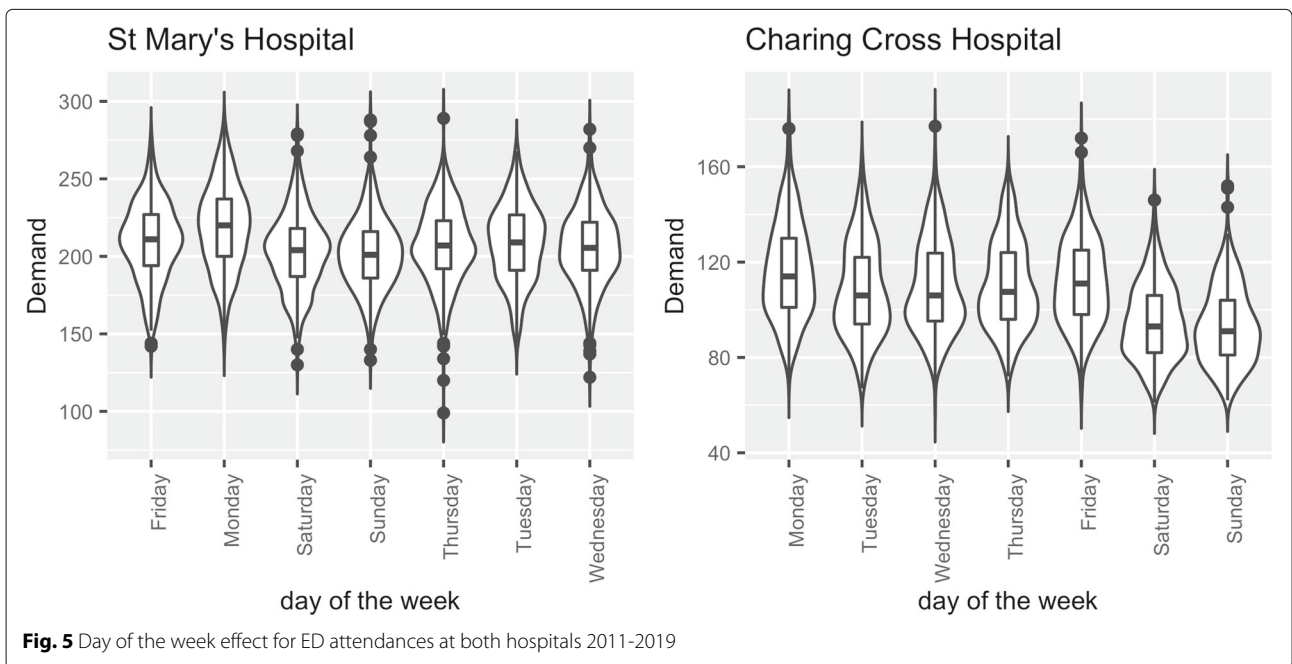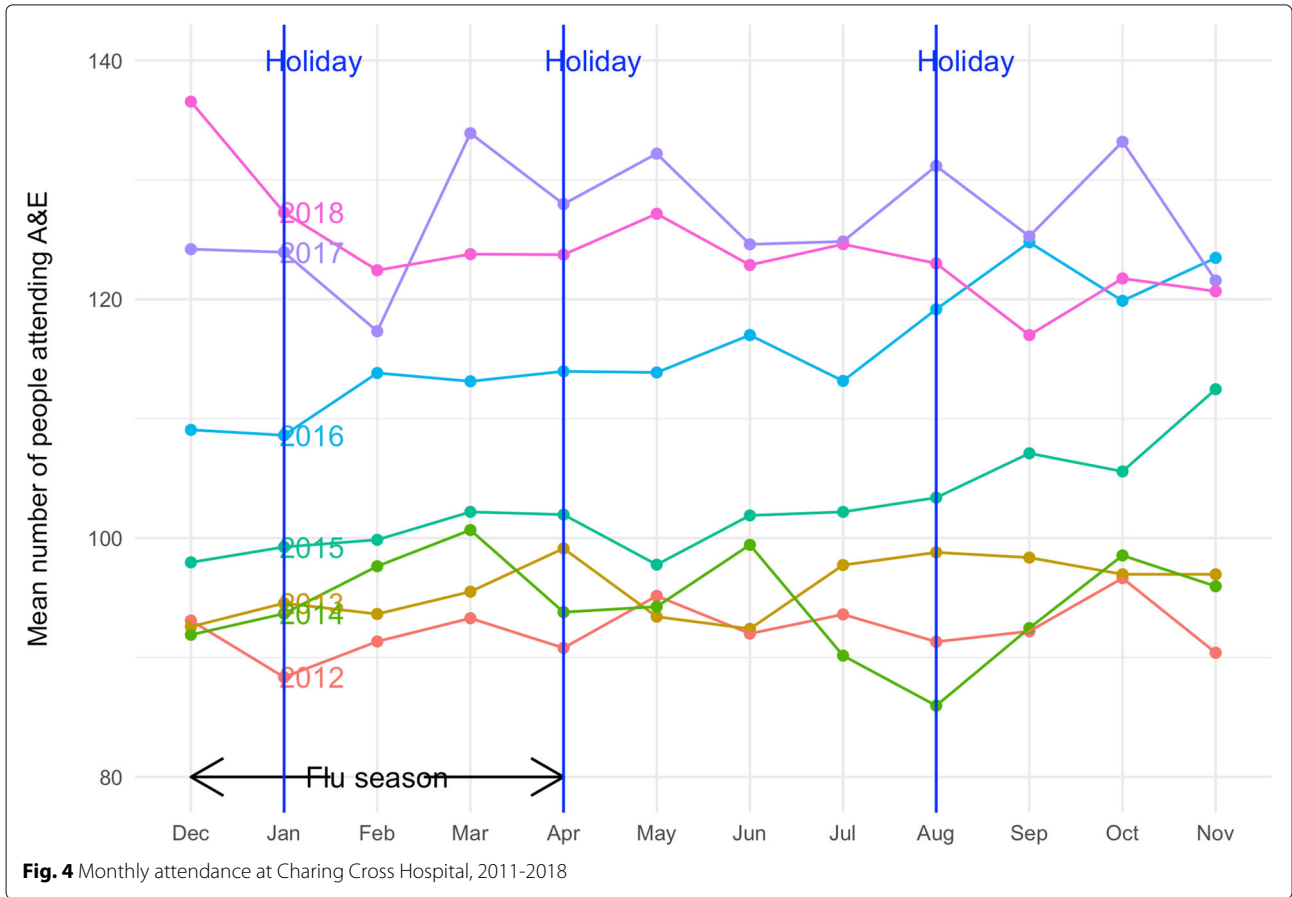
Fig. 4 Monthly attendance at Charing Cross Hospital, 2011-2018

Fig. 5 Day of the week effect for ED attendances at both hospitals 2011-2019

Table 1 Excerpt from the model matrix corresponding to St Mary's Hospital. Explanatory covariates for the machine learning algorithms are listed

Date	2014-04-10	2014-04-11	2014-04-12	2014-04-13	...
demand	198	183	172	185	...
month	April	April	April	April	...
yesterday	218	198	183	172	...
same day last week	223	189	187	195	...
average of previous week	199.6	196	195.1	193	...
time	−0.306	−0.305	−0.304	−0.303	...
bank holiday	FALSE	FALSE	FALSE	FALSE	...
school holiday	FALSE	FALSE	TRUE	TRUE	...
day of week	Thursday	Friday	Saturday	Sunday	...
precipitation on previous day	0	0	0	0	...
max temperature on previous day	17.5	17.2	15.0	17.4	...
min temperature on previous day	12.0	11.8	10.8	11.3	...
flu hits on google on previous day	7.52	4.88	5.12	4.08	...
Notting Hill carnival	FALSE	FALSE	FALSE	FALSE	...
Christmas	FALSE	FALSE	FALSE	FALSE	...

training window size remains fixed. We employ the latter method.

Using temporal cross-validation for time series algorithms only requires splitting the data into a training and test set. Adapting the data so that the last 730 days (24 months) of data are held out for testing, 2140 days (roughly six years) were available for training which corresponds to a ~75%/25% split. We then applied temporal cross validation to each day in the test set (see Fig. 6) using a rolling window so that all training sets consist of the same number of days.

Given the fairness and robustness of our cross validation scheme, we are confident that our results are robust to data shift and not only valid for the data times we have collected.

Validation of hyperparameters
In order to allow for a fair comparison of the machine learning algorithms, the method for temporal cross-validation has to be adapted as most machine learning algorithms require tuning of their hyperparameters to balance over and under fitting. Therefore we also have to split the training data into a training and a validation set in order to choose the best set of hyperparameters which minimizes the error on the validation set. All machine learning algorithms listed above (see "Machine learning
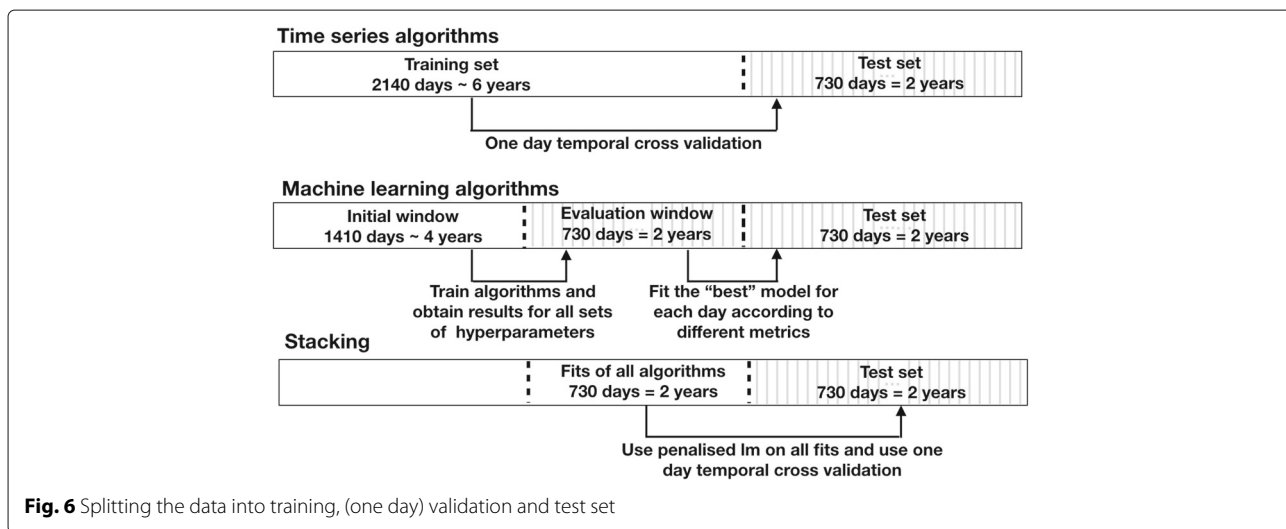
Fig. 6 Splitting the data into training, (one day) validation and test set

algorithms" section) were trained on 4 years worth of data and their hyperparameters were compared on a validation set consisting of 730 days. In our analysis we have used two different approaches to split the training set and to choose the hyperparameters. We will call these the batch and the online method.

The batch method

In order to choose the set of hyperparameters which minimizes the error on the validation set, we consider five different approaches. We choose the set of hyperparameters which minimize the error

1. on the previous day,
2. on the past *n* days,
3. over the whole validation period using an exponential moving average,
4. averaged over the whole validation period, or
5. according to caret's built in rules (see [30]).

For each of the five cases above, we choose the best set of hyperparameters for each day of the test set and refitted all models on a daily basis. Of course refitting every model for each day of the test set is computationally expensive. Therefore we develop an online method, described below.

The online method

In the batch method described above, all models are refitted daily, which takes significant computational power to run and might not be feasible for a deployed version of our methods in a hospital setting. Thus, we explore whether keeping the parameters fixed for longer periods, which yields significant savings in terms of computation, hurts performance. We refit each model over several testing periods of various sizes which are subsets of the test set of

length 730 days with a rolling origin, see Fig. 7. Our chosen testing periods are 1 day, 7 days, 30 days, 60 days, 365 days and 730 days long. That means that we validated the best set of hyperparameters for each algorithm for every testing period, and then rolled forwards. The final error rates are the result of the overall error on all predictions on the whole test set of length 730.

Results

We compare the mean absolute error (MAE) and mean absolute percentage error (MAPE) rates for all time series algorithms as well as for the batch and the online method as outlined above (see "The batch method" and "The online method" sections). Finally, we compare our results with the stacked regression.

St Mary's hospital

The MAE of the time series algorithms range from 14.46 to 17 as shown in Table 2, with an MAPE ranging from 6.9% to 8.3%. Hence, although the time series algorithms are simpler and only based on the time series without using any other predictors, such as weather, bank or school holidays (see Table 1 for details on all predictors), they already give relatively accurate results. In case of the machine learning algorithms, independent of the type of hyperparameter tuning we use, most of the results for the batch methods range between an MAE of 14.3 to 15.2 and an MAPE of 6.8% − 7.4%, as shown in Table 3. Only the k-nearest neighbour algorithm performs worse despite different ways of tuning. Especially when using the online method, see Table 4a, the linear models produce the best results for St Mary's hospital with an MAPE of 6.8% in the case of daily retraining of the generalized linear model.
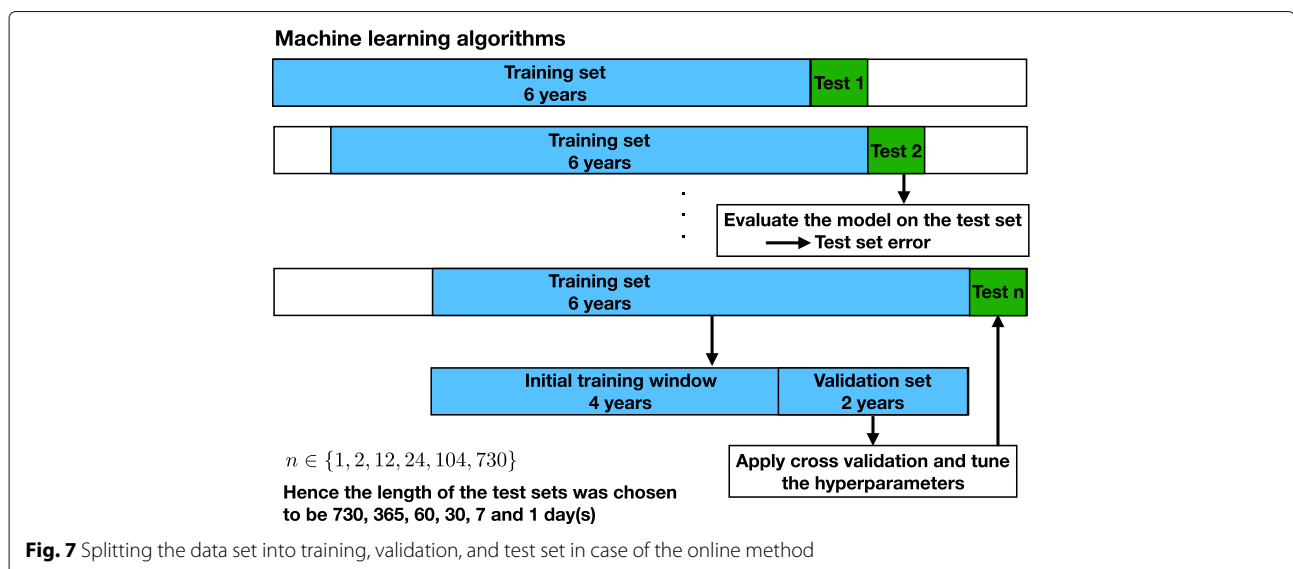

Fig. 7 Splitting the data set into training, validation, and test set in case of the online method

Table 2 Error rates for all time series algorithms for St Mary's and Charing Cross Hospital

| algorithm | days | St Mary's Hospital: | | Charing Cross Hospital: | |
		MAE	MAPE (in %)	MAE	MAPE (in %)
	1	15.29	7.2	12.33	10.1
ARIMA	3	15.68	7.4	14.75	12.2
	7	15.58	7.4	13.81	11.4
	1	14.47	6.9	10.46	8.5
ETS	3	16.89	8.0	17.60	14.6
	7	16.18	7.6	14.55	11.8
	1	14.54	6.9	10.72	8.7
StructTS	3	17.35	8.2	17.86	14.8
	7	16.54	7.8	14.72	12.0
	1	14.51	6.9	10.59	8.6
STLM	3	16.67	7.9	16.83	13.9
	7	16.00	7.6	14.28	11.6

Charing cross hospital

The results for Charing Cross Hospital are similar although the error rates are slightly increased. As shown in Table 2, most time series algorithms yield MAE error rates between 10.5% and 14.5% while the corresponding MAPE error rates range from 8.5% to 12%. The best performance reached using the batch method is also around 10.5% although the overall performance is a little bit better, see Table 5. In case of the online method, the MAPE error rates are as low as 8.6% for gradient boosting machine, the generalized linear model or a simple linear model retrained on a daily basis, as shown in Table 4b.

Stacked regression

In order to make use of the strengths of all algorithms, we applied a generalised linear model as well as a penalized regression to create an effective ensemble approach, see Fig. 6 for details. The best performance was achieved

Table 3 MAE rates for all types of hyperparameter tuning predicting hospital demand 1, 3 or 7 days in advance in case of the batch method for St Mary's hospital

| algorithm | days | Choosing the best set of hyperparameters based on: | | | | |
		yesterday	the past *n* days	exponential moving average	the average over the whole training set	caret
	1	14.50	14.38	14.48	14.42	14.34
lm	3	14.80	14.96	14.86	14.95	14.80
	7	15.11	15.17	15.22	15.33	15.13
	1	15.97	14.49	15.08	14.46	14.30
gbm	3	15.38	14.63	14.86	14.67	14.53
	7	15.43	14.79	14.93	14.78	14.86
	1	14.49	14.34	14.38	14.35	14.48
glmnet	3	14.81	14.79	14.82	14.75	14.77
	7	14.98	15.11	14.94	15.16	15.09
	1	15.86	15.47	15.74	15.79	15.42
knn	3	16.13	15.78	15.92	16.02	15.57
	7	16.70	16.05	16.45	16.52	15.89
	1	14.62	14.46	14.54	14.79	14.45
rf	3	14.66	14.55	14.54	15.15	14.53
	7	14.93	14.61	14.62	15.53	14.67

Table 4 Results from the online method for both hospitals

algorithm	period	MAE	MAPE (in%)	algorithm	period	MAE	MAPE (in %)
lm	1	14.27	6.7	lm	730	10.59	8.6
glmnet	1	14.31	6.8	lm	1	10.59	8.6
gbm	730	14.33	6.8	glmnet	1	10.60	8.6
lm	7	14.33	6.8	lm	7	10.62	8.6
lm	365	14.34	6.8	gbm	1	10.63	8.6
glmnet	7	14.37	6.8	glmnet	7	10.64	8.6
glmnet	365	14.38	6.8	glmnet	730	10.64	8.5
gbm	7	14.40	6.8	gbm	7	10.71	8.7
lm	60	14.47	6.8	gbm	730	10.78	8.7
glmnet	60	14.47	6.8	lm	365	10.80	8.7
glmnet	1	14.49	6.9	glmnet	365	10.82	8.7
lm	30	14.50	6.8	lm	30	10.84	8.8
glmnet	30	14.50	6.8	glmnet	30	10.84	8.8
gbm	30	14.52	6.9	lm	60	10.86	8.8
gbm	60	14.52	6.9	glmnet	60	10.87	8.8
gbm	365	14.53	6.9	rf	1	10.93	8.9
rf	1	14.55	6.9	gbm	365	10.96	8.9
glmnet	730	14.58	6.8	gbm	30	11.00	8.9
lm	730	14.60	6.9	rf	7	11.08	9.0
rf	7	14.66	6.9	gbm	60	11.15	9.0
rf	730	14.73	6.9	rf	60	11.21	9.1
rf	365	14.76	7.0	rf	30	11.21	9.1
rf	60	15.02	7.1	rf	365	11.49	9.2
rf	30	15.08	7.1	rf	730	11.51	9.0
knn	1	15.52	7.3	knn	7	12.55	10.1
knn	7	15.53	7.3	knn	1	12.60	10.1
knn	730	15.61	7.4	knn	30	12.75	10.2
knn	365	15.62	7.4	knn	60	12.78	10.2
knn	30	15.75	7.4	knn	730	12.81	10.1
knn	60	15.93	7.5	knn	365	12.81	10.1
	(a) St Mary's Hospital				(b) Charing Cross Hospital		

using penalized regression with an MAE error rate of 14.51 for St Mary's Hospital and 10.60 for Charing Cross Hospital.

Interpretation of results

In our analysis we consider a variety of predictors ranging from past values of ED attendance, to the weather forecast and to school and bank holidays. The question left to answer is therefore which of the predictors are actually important for making accurate predictions? Could some of them actually be redundant?

In machine learning, variable importance can be defined as the dependence between input and output variables and computed by permuting the values of a given predictor and calculating error on a held out set. This measure has drawbacks in the case of multicollinearity, which could suppress the importance of certain variables in some models. As shown in Fig. 8, the most important variable predicting demand is the average demand the week before - except for glmnet - followed by specific days and months. The importance of these variables varies considerably between algorithms but largely is consistent with one another. To our knowledge, this is the first attempt to quantify the importance of variables in predicting demand. We believe that these importance percentages can be utilised as heuristics to help staff improve their prediction of demand in the absence of statistical analysis.

Table 5 MAE error rates for all types of hyperparameter tuning predicting hospital demand 1, 3 or 7 days in advance in case of the batch method for Charing Cross hospital

| algorithm | days | Choosing the best set of hyperparameters based on: | | | | |
		yesterday	the past *n* days	exponential moving average	the average over the whole training set	caret
	1	10.66	10.60	10.67	10.58	10.51
lm	3	10.75	10.72	10.84	10.69	10.69
	7	10.78	10.90	10.83	10.87	10.77
	1	12.50	10.68	11.12	10.68	10.89
gbm	3	12.49	10.58	10.92	10.50	10.78
	7	12.11	10.73	11.03	10.57	10.72
	1	10.72	10.51	10.59	10.51	10.51
glmnet	3	10.85	10.69	10.74	10.69	10.69
	7	10.86	10.76	10.84	10.76	10.76
	1	12.81	12.57	12.62	12.77	12.36
knn	3	12.85	12.58	12.78	12.87	12.51
	7	12.65	12.45	12.41	12.47	12.33
	1	11.09	10.84	11.04	11.55	10.89
rf	3	10.98	10.80	10.83	11.85	10.78
	7	10.92	10.78	10.79	11.99	10.72

Discussion

The results of our analysis highlight that in our case a penalised linear method performs equally well or better when compared to more sophisticated machine learning methods such as the gradient boosting algorithm, the random forest algorithm or the stacked regression. However, each of these methods has desirable advantages over the others. While a linear model is much easier to apply in practice, machine learning algorithms can generalize better out of sample. Combining both the machine learning and linear modelling approaches improve the diversity of model predictions so that stacked predictions are more robust and accurate than any single model including the best performing one. This has been formally demonstrated [31] but is also intuitive - it follows from a 'wisdom of the crowd' rule for ensemble learning. However, using a stacked regression comes with the disadvantage that the valuable insight on the drivers of a time series may get lost. Hence the mechanism presented above can be understood as a road map to find the best method for a particular problem weighing its accuracy, simplicity and interpretability. If additional covariates are collected for our problem at hand or the data is more granular, e.g. hourly, this process needs to be reapplied.

The online method that we have created has the ability to provide accurate results with a very quick turnaround time: an average model only takes a couple of minutes to train and forecast. Running it over longer periods of time without retraining of the model is much faster and at the same time not significantly worse than tuning the hyperparameters on a daily basis. These properties are highly desirable from an operational standpoint. Our model is easily deployed to perform near real time analysis to help staff and health care planners adjust existing rotas or prepare for excess demand, especially important when health systems are stressed by acute health events such as COVID-19 [32].

Conclusion

The framework, analysis and methodology proposed in this paper are highly relevant from an operational viewpoint. To the authors knowledge, the majority of EDs in the UK do try to account for demand, but do so using ad hoc heuristics. Our approach provides some scientifically backed information to improve these heuristics, but more importantly provides a framework that is quick and easy to implement. Estimates of 1, 3 and 7 day demand forecasts can be created at any time and updated easily allowing statistically backed estimates to be used to inform hospital policy and practice. Our hope is this paper will fill a knowledge gap and increase hospitals uptake in the use of these methods. Given epidemics and disease outbreaks that can strain health systems, our approach can provide added precision to help EDs operate as efficiently as possible. The challenge will then be for teams in hospitals to implement more insight-driven ways of working,
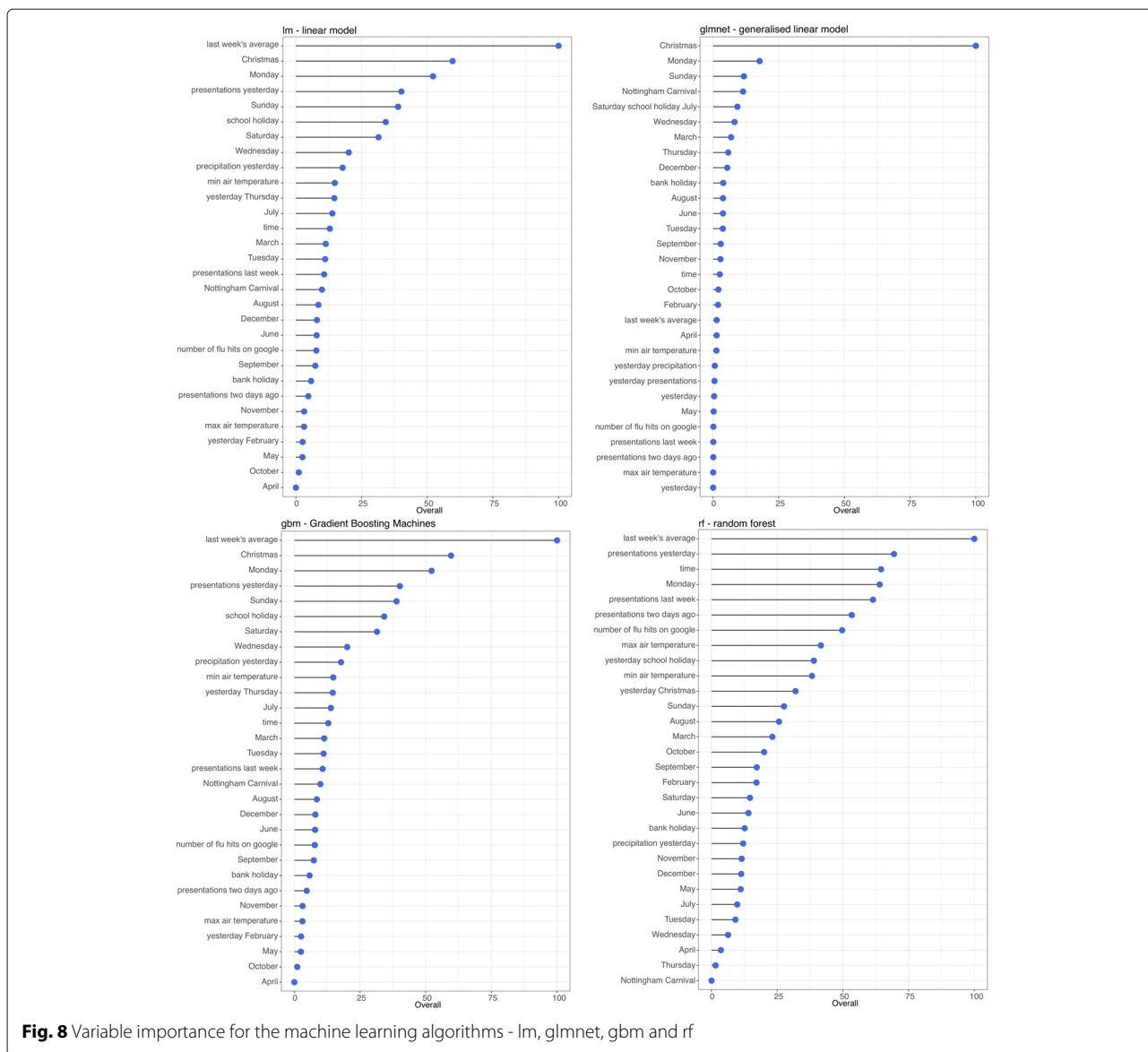
Fig. 8 Variable importance for the machine learning algorithms - lm, glmnet, gbm and rf

more flexible approaches to rostering staff and innovative ways of communicating with their local com munities.

In an era of precision medicine, future work will undoubtedly focus on granular data sets including individual level patient data with diagnoses and testing information. This will not only help to improve performance of the predictions but also help to understand dependencies between seasonality, events and reasons for ED presentations.

Abbreviations
ED: Emergency department; NHS: The national health service

Authors' contributions
MV, SF and SB conceived the study. MV, SF and SB wrote the first manuscript draft. MV, SF and SB analysed the data. MV, SF and SB interpreted the data. BG

and LC prepared the data set. GC and RK provided public health relevance expertise. All authors reviewed, edited and substantially revised the manuscript for scientific content. All author(s) read and approved the final manuscript.

Availability of data and materials
The data sets analysed during the current study have been provided by Imperial College NHS Trust. They have been generated from individual patient

level data sets but have not been made publicly available. However,they are available on reasonable request. Please contact Dr Samir Bhatt for details.

Author details
[1]MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, UK. [2]Imperial College Healthcare NHS Trust, London, UK. [3]Department of Mathematics and Data Science Institute, Imperial College London, London, UK. [4]Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK.

References

1. Baker C. NHS Key Statistics, England, February 2020. 2020. https://commonslibrary.parliament.uk/research-briefings/cbp-7281/.
2. Baker C, House of Commons Library. NHS key statistics: England, May 2019. 2019. https://commonslibrary.parliament.uk/research-briefings/cbp-7281/. Accessed 1 Jan 2021.
3. McCarthy ML. Overcrowding in emergency departments and adverse outcomes. BMJ. 2011;342:d2830.
4. Silvester K, Lendon R, Bevan H, Steyn R, Walley P. Reducing waiting times in the NHS: is lack of capacity the problem? Clin Manag. 2004;12(3):105–9.
5. NHS England. Reference costs 2017/18: highlights, analysis and introduction to the data. 2018. https://improvement.nhs.uk/documents/1972/1_-_Reference_costs_201718.pdf. Accessed 1 Jan 2021.
6. Acute Care and Workforce/ Workforce/Pay, Pensions & Employment Services /13710. The NHS Pay Review Body (NHSPRB) Review for 2018: Written Evidence from the Department for Health and Social Care for England. 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/676185/NHSPRB_evidence_for_OME_PDF1.pdf. Accessed 1 Jan 2021.
7. Buchan J, Gershlick B, Charlesworth A, Seccombe I. Falling short: the NHS workforce challenge: The Health Foundation; 2019. https://www.health.org.uk/sites/default/files/upload/publications/2019/S05_Falling%20short_The%20NHS%20workforce%20challenge.pdf.
8. Boyle J, Jessup M, Crilly J, Green D, Lind J, Wallis M, Miller P, Fitzgerald G. Predicting emergency department admissions. Emerg Med J. 2012;29(5):358–65.
9. McCarthy ML, Zeger SL, Ding R, Aronsky D, Hoot NR, Kelen GD. The challenge of predicting demand for emergency department services. Acad Emerg Med. 2008;15(4):337–46.
10. Jones SS, Thomas A, Evans RS, Welch SJ, Haug PJ, Snow GL. Forecasting daily patient volumes in the emergency department. Acad Emerg Med. 2008;15(2):159–70.
11. Champion R, Kinsman LD, Lee GA, Masman KA, May EA, Mills TM, Taylor MD, Thomas PR, Williams RJ. Forecasting emergency department presentations. Aust Health Rev. 2007;31(1):83–90.
12. Hoot NR, LeBlanc LJ, Jones I, Levin SR, Zhou C, Gadd CS, Aronsky D. Forecasting emergency department crowding: a discrete event simulation. Ann Emerg Med. 2008;52(2):116–25.
13. City of Westminster. School term and holiday dates. 2019. https://www.westminster.gov.uk/school-term-and-holiday-dates. Accessed 1 Jan 2021.
14. gov.uk. UK bank holidays. 2019;https://www.gov.uk/bank-holidays:.
15. B. F. Google Trends: How to acquire daily data for broad time frames. 2019. https://medium.com/@bewerunge.franz/google-trends-how-to-acquire-daily-data-for-broad-time-frames-b6c6dfe200e6. Accessed 1 Jan 2021.
16. Google. Google Trends. 2019. https://trends.google.com/trends.
17. The Health and Social Care Information Centre. Accident and Emergency Attendances in England - 2008-2009, Experimental statistics. 2010. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-accident--emergency-activity/accident-and-emergency-attendances-in-england-2008-2009-experimental-statistics.
18. NHS England and NHS Digital. Hospital Accident and Emergency Activity 2017-18. 2018. https://files.digital.nhs.uk/D3/CCB4FE/AE1718_%20Annual%20Summary.pdf.
19. Makridakis S, Spiliotis E, Assimakopoulos V. The M4 Competition: 100,000 time series and 61 forecasting methods. Int J Forecast. 2020;36(1):54–74.
20. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457(7232):1012.
21. Hyndman RJ, Athanasopoulos G. Forecasting: Principles and Practice, 2nd edn. Australia: OTexts; 2018.
22. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22. http://www.jstatsoft.org/v33/i01/.
23. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017;77(1):1–17. https://doi.org/10.18637/jss.v077.i01.
24. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.
25. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189–232. https://doi.org/10.1214/aos/1013203451.
26. Bishop CM, Vol. 4. Pattern Recognition and Machine Learning; 2006, p. 738. https://doi.org/10.1117/1.2819119. http://www.library.wisc.edu/selectedtocs/bg0137.pdf.
27. Breiman L. Stacked regressions. Mach Learn. 1996;24(1):49–64.
28. Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241–59.
29. Bhatt S, Cameron E, Flaxman SR, Weiss DJ, Smith DL, Gething PW. Improved prediction accuracy for disease risk mapping using Gaussian Process stacked generalisation. J R Soc Interface. 14(134):20170520. https://doi.org/10.1098/rsif.2017.0520.
30. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28:. https://doi.org/10.18637/jss.v028.i05.
31. Krogh A, Vedelsby J. Neural network ensembles, cross validation and active learning. In: Proceedings of the 7th International Conference on Neural Information Processing Systems, NIPS'94. Cambridge,: MIT Press; 1994. p. 231–8.
32. Vollmer M, Radhakrishnan S, Kont M, Flaxman S, Bhatt S, Costelloe C, Honeyford C, Aylin P, Cooke G, Redhead J, et al. Report 29: The impact of the COVID-19 epidemic on all-cause attendances to emergency departments in two large London hospitals: an observational study. 2020. https://spiral.imperial.ac.uk/bitstream/10044/1/80295/12/2020-07-01-COVID19-Report-29.pdf. Accessed 1 Jan 2021.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.