# Probabilistic Forecasting of Patient Waiting Times in an Emergency Department

Siddharth Arora[1], James W. Taylor[1] and Ho-Yin Mak[2]

[1]*Saïd Business School, University of Oxford, U.K.*

[2]*McDonough School of Business, Georgetown University, U.S.*

*Manufacturing & Service Operations Management*, forthcoming.

**Problem definition**: We study the estimation of the probability distribution of individual patient waiting times in an emergency department (ED). While it is known that waiting time estimates can help improve patients' overall satisfaction and prevent abandonment, existing methods focus on point forecasts, thereby completely ignoring the underlying uncertainty. Communicating only a point forecast to patients can be uninformative and potentially misleading.

**Methodology/Results**: We use the machine learning approach of quantile regression forest (QRF) to produce probabilistic forecasts. Using a large patient-level dataset, we extract the following categories of predictor variables: (1) calendar effects, (2) demographics, (3) staff count, (4) ED workload due to patient volumes, and (5) the severity of the patient condition. Our feature-rich modelling allows for dynamic updating and refinement of waiting time estimates as patient- and ED-specific information (e.g., patient condition, ED congestion levels) is revealed during the waiting process. The proposed approach generates more accurate probabilistic and point forecasts, when compared with methods proposed in the literature for modelling waiting times, and rolling average benchmarks typically used in practice.

**Managerial Implications**: By providing personalized probabilistic forecasts, our approach gives low-acuity patients and first responders a more comprehensive picture of the possible waiting trajectory, and provides more reliable inputs to inform prescriptive modelling of ED operations. We demonstrate that publishing probabilistic waiting time estimates can inform patients and ambulance staff in selecting an ED from a network of EDs, which can lead to a more uniform spread of patient load across the network. Aspects relating to communicating forecast uncertainty to patients, and implementing this methodology in practice, are also discussed. For emergency healthcare service providers, probabilistic waiting time estimates could assist in ambulance routing, staff allocation, and managing patient-flow, which could facilitate efficient operations and cost savings while aiding better patient care and outcomes.

*Key words:* low-acuity; machine learning; managing patient-flow; routing, quantile regression forest.

---

## 1. Introduction

Emergency departments (EDs) are coming under increasing pressure to provide safe and quality care to patients in a timely manner. It was estimated that between 2006 and 2016, there were 1.4 billion ED visits in the US alone, whereby the number of visits increased by around 2.3 million per year (Singer et al. 2019). The hospital staffing and infrastructure has not grown at the same rate, which has resulted in longer waiting times. In England, the national health service (NHS) has a pledge set out in the handbook of its constitution stating that 95% of patients attending the ED should be treated, admitted or discharged within four-hours (NHS 2019). However, for the first time in 2019, all major ED units in England missed their waiting time targets; while in the US, from 2003 to 2009, the mean waiting time increased by 25% (Hing and Bhuiya 2012). This is a matter of growing concern, as long ED waiting times are associated with increased morbidity and mortality, and are one of the leading causes of patient dissatisfaction (Bernstein et al. 2009). Waiting times and perceived queue length can influence patients to dropout from the ED (Batt and Terwiesch 2015). Patients that dropout are linked with having a higher likelihood of re-presentation and poor outcomes (Carter et al. 2014). Providing delay estimates can help lower the perceived waiting time (Jouini et al. 2011).

Numerous studies have focused on forecasting demand at the ED, such as attendances, ambulance arrivals, and admissions; however, the literature on modelling personalized waiting times is relatively scarce. Note that demand at the ED is recorded at equally spaced time intervals, typically each hour or day, and so time series models are an obvious choice when forecasting demand. Waiting times, on the other hand, are recorded for each patient, which makes patient-level cross-sectional modelling a more natural choice. The methods used for modelling demand cannot be directly translated for forecasting waiting times, which makes these two tasks conceptually different in terms of their potential use cases and the underlying modelling framework. Our aim is to contribute to the literature on waiting times using a probabilistic modelling approach. Future estimates of demand are necessary for capacity planning and resource allocation, while accurate estimates of waiting times can help patients select an ED site from a network of EDs, which can help streamline patient flow and mitigate overcrowding. Growing the capacity in EDs to eradicate congestion and minimize waiting times requires long-term planning and funding; meanwhile, providing patients with an estimate of their waiting times can be an inexpensive and immediate way forward to managing patient expectations and reducing abandonment rates, thereby improving patient outcomes and quality of care.

For EDs, long waiting times can have significant economic implications. In private healthcare systems (e.g., in the US), shorter and more transparent waiting times could lead to higher revenues for hospitals, since about 10% of total US healthcare cost is spent on emergency care (Galarraga and Pines

2016). In many OECD countries with public healthcare systems, service providers can incur financial penalties if they exceed the waiting time targets (Summary 2013). It is estimated that prolonging waiting time in the ED by just 10 minutes increases the cost of care by an average of 6% for a high-acuity patient, and 3% for a low-acuity patient (Woodworth and Holmes 2020). To deal with long waiting times (or congestion), EDs sometimes rely on external agencies to provide temporary workforce. However, temporary staff cost 20% more on average than the permanent staff (Buchan et al. 2019). Waiting time estimates could potentially assist hospitals in making more informed staffing decisions, thereby reducing the dependency on costly surge capacities. Given the impact of long waiting times on patient outcomes and its economic implications for service providers; streamlining patient-flow, minimizing waiting times, and optimizing resource allocation, while providing quality care, is at the heart of reforming services offered by the EDs.

Modelling of ED operations has been an active area of research in operations management. A significant stream of literature takes descriptive and prescriptive views on patient flows in the ED by use of queueing models (Armony 2015; Batt and Terwiesch 2015; Bayati 2017; Xu and Chan 2016), as well as discrete-event simulations (Baril et al. 2019). For a detailed review of literature in this area, see, for example the papers by Hu et al. (2018); Keskinocak and Savva (2020); Misic and Perakis (2020); Singh and Terwiesch (2012), and references therein.

Accurate predictive models for waiting times can provide valuable information to patients, assist service providers in planning and operations, as well as support prescriptive studies of ED operations (e.g., for calibrating queueing and simulation models). Ang et al. (2016) generate point forecasts for ED waiting times (from registration to start-of-treatment) based on the least absolute shrinkage operator (Lasso) using predictor variables inspired by fluid model estimators. They report a reduction of over 30% in mean squared error, compared to a rolling average model, which is the standard method adopted in practice in US hospitals (Dong et al. 2019). Ding et al. (2010) generate estimates of treatment, boarding, and waiting room time using quantile regression, focussing on 10, 50, and 90% quantiles. Using data available at triage, Sun et al. (2012) use quantile regression to predict the median and 95% quantile of the waiting time (from triage to consultation). We are not aware of any existing study that models and evaluates the whole probability distribution of ED waiting times.

Over the past decade or so, the forecasting literature has moved beyond point forecasts to emphasize the importance of conveying forecast uncertainty through probabilistic forecasts (Gneiting and Katzfuss 2014). The application and context drive the choice of probabilistic forecasting method, with data availability and forecast horizon being important issues. Parametric univariate models are often used if the only information available is the historical time series (see, for example, Taylor 2012). When a

group of time series must be predicted, capturing interdependencies can be important (see, for example, Ye 2019). If some series in a group equal the sum of others, as in a hierarchy, constraints must be imposed when generating probabilistic forecasts (see, for example, Ben Taieb et al. 2021). Sometimes, the only forecasts available are judgemental (see, for example, Gaur et al. 2007). If a rich set of variables are available, models such as multiple linear regression could be used. However, if linearity and a distributional assumption are inappropriate, and a large dataset is available, nonparametric machine learning methods offer a promising alternative. Recent examples of this are provided by Guo et al. (2021), who use a regression tree to produce probabilistic forecasts of individual airport passenger connection times, and by Salari et al. (2022), who use an ensemble of such trees in an application of the quantile regression forests (QRF) of Meinshausen (2006) to generate probabilistic forecasts of delivery times in online retailing. In a similar vein, we also use a QRF for the probabilistic forecasting of durations, with our focus being ED patient waiting times.

Our proposed approach differs from the previous studies of forecasting ED waiting times in that it is: (1) probabilistic, and (2) personalized. While accurate forecasting of waiting times is the main objective of this study, we also explore the operational implications of this approach by investigating the following questions:

(a) Is there value in using point estimates of ED waiting times for making routing decisions, as compared to using, say, either the distance or travel times to the ED?
(b) Is there value in using probabilistic estimates of ED waiting times for making routing decisions compared to using only a point estimate of waiting time?
(c) Is there value in providing waiting time estimates to both patients and EMTs, compared to providing such estimates to only one of them?

We show that personalized and probabilistic estimates of waiting times can assist patients and emergency medical technicians (EMTs) make informed routing decisions while selecting an ED, which in turn results in a uniform spread of patient load and lower congestion across EDs in a geographic neighbourhood. It can be surmised that high waiting times (from registration to initial assessment) would subsequently result in long patient length-of-stay (LOS). Using historical data, we show the relation between waiting times and the probability of a patient breaching the 4-hour LOS target, which can be used by EDs to help prioritize patients at the time of registration. Waiting time is inherently uncertain and the distribution is asymmetric, and so a point forecast can be uninformative and even potentially misleading, which could risk greater dissatisfaction among patients. Given that patients can feel increasingly dissatisfied if they end up waiting longer than the published estimate, it is imperative that the uncertainty associated with forecasts is adequately conveyed, so that patient expectations can be better managed. When only a deterministic forecast is provided, people tend to make assumptions

about the forecast uncertainty (Morss et al. 2008). This underscores the importance of communicating the uncertainty to the patients. In this study, we generate and evaluate probabilistic forecasts, rather than focussing on just a point estimate or predefined quantiles. Patient waiting times depend on a multitude of complex *features* (or predictor variables), some of which are also time-varying (such as, staff count, queue length) and/or patient-specific. Incorporating such individual and time-varying features enables us to produce personalized forecasts that are updated over time, as new information regarding the ED's utilization and the patient's conditions are observed. The potential nonlinear relationship between high-dimensional features that help characterize patient-flow in the ED and waiting times motivates our use of machine learning. Specifically, we employ a QRF to estimate conditional probability distributions of the waiting times in a nonlinear and nonparametric framework Moreover, to investigate key predictors of ED waiting times, we use a random forest (RF) to derive rankings of feature importance. This information could provide useful insights into patient flow and potential bottlenecks in the ED.

Using a large patient-level dataset, we extract five different categories of features, that quantify: (1) calendar effects (time of arrival), (2) demographics (age, sex), (3) staff count, (4) ED workload due to patient volumes (attendances), and (5) the severity of patient condition. We provide a comparison of the QRF-based approach with Q-Lasso (Ang et al. 2016), quantile regression (Ding et al. 2010; Sun et al. 2012), *k*-nearest neighbour, and rolling average benchmarks that are typically used in practice. Model evaluation is based on a comparison of distributional, quantile, and point forecast accuracy.

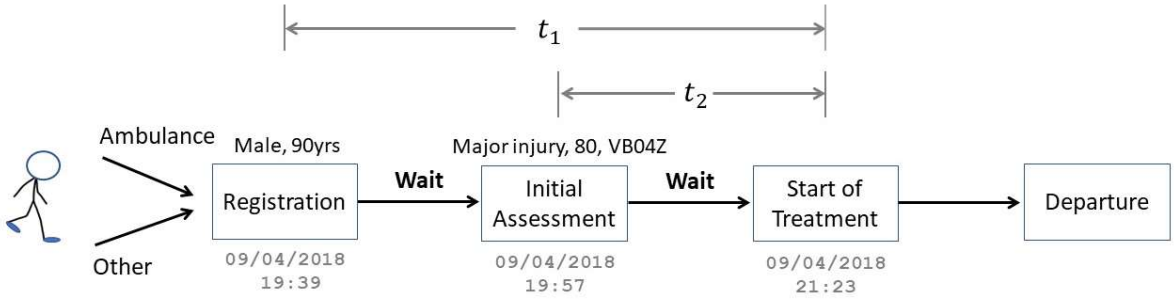
We further discuss three aspects related to the practical implementation of the proposed methodology. Firstly, we propose and evaluate a colour-coded (categorical) scheme to communicate forecast uncertainty to the patients. Secondly, we show that our modelling scheme could be used to provide updates of waiting time estimates, by incorporating additional information regarding triage, which becomes available at the time of initial assessment. Finally, we demonstrate that personalized probability distribution estimates of waiting times, when used in conjunction with travel time estimates, could help patients select an ED site from a geographic network of EDs, and could result in a more uniform spread of workload among EDs. Patients could access waiting and travel time estimates, on say, a smartphone application. The findings of this study could have direct implications for both patients and EDs. The paper is arranged as follows. Section 2 describes patient-flow and presents our ED dataset. Section 3 presents features and the quantile regression forest. Section 4 presents naïve and sophisticated benchmark methods. Empirical comparison of forecast accuracy is provided in Section 5. Illustrative examples demonstrating the practical applications of the modelling framework are presented in Section 6. Section 7 summarizes the paper and discusses future work.

## 2. ED Patient-flow and Data

### 2.1 Understanding patient-flow in the ED

This study employs data from the John Radcliffe (JR) hospital in Oxford, UK. This hospital treats both minor and major injuries, and is operational 24 hours a day. The policies at the JR are determined by the NHS, and hence the operations at this hospital are similar to other hospitals in the UK. Figure 1 presents a basic schematic diagram of ED patient flow at the JR hospital. Patients arrive at the ED either by ambulance or another mode of transport. Upon arrival, patients register at the reception desk, where they provide the following information: name and address, date of birth, reason for visit, name of the general practitioner (doctor) with whom the patient is registered (we denote the time of registration as $t_{reg}$). After registration, patients are requested to take a seat in the ED waiting room until they are called for an initial assessment. At the time of initial assessment (denoted by $t_{assess}$), patients are categorized by a nurse using a: (1) patient group number (code denoting reason for presenting complaint), (2) human resource group (code denoting use of resources), and (3) triage category (to prioritize patients depending upon their severity, the different triage categories are 'minor injury', 'major injury', 'urgent care', or 'resuscitation'). Patients who are triaged as 'minor injury' might need to wait longer in the queue, compared to other patients who have a more serious health condition. Patients with critical medical needs are triaged as 'urgent care' or 'resuscitation', and are seen with priority by a doctor upon arrival at the ED. Thus, in this study, we only generate waiting time estimates for patients that are triaged as either 'minor injury' or 'major injury' (we refer to these as *low-acuity* patients). Following the initial assessment, a patient returns to the waiting area until they are called by a nurse to start treatment. At the time of treatment (denoted by $t_{treat}$), the patient is seen by a doctor. Depending upon the outcome of the treatment, the patient departs from the ED (patients either leave the hospital, or get admitted to the ICU or another ward within the hospital).

**Figure 1.** Schematic diagram of typical patient-flow in the ED.




*Note*: Figure presents timestamps of registration, initial assessment and treatment (for a 90-year-old male, triaged as 'major injury', patient group '80' and code 'VB04Z'). $t_1$ and $t_2$ are 104 and 86 minutes, respectively. Using information available at the time of registration, we predict $t_1$. To predict $t_2$, we use additional information about patient symptoms available at the time of initial assessment.

Studies have suggested that patients are more sensitive to their start-of-treatment time than their time-of-departure (Boudreaux et al. 2000). Once with the physician, the patient is typically far more tolerant of the passage of time (Anderson et al. 2007). This motivates us to focus on the time spent waiting until the start of treatment. In addition to receiving, when they arrive, an estimate of their waiting time to treatment, it is helpful to receive updates to this during their wait. We thus update the waiting time estimate for each patient. Specifically, for each low-acuity patient, we model two waiting times: (1) $t_1$: time from registration to start-of-treatment ($t_1 = t_{treat} - t_{reg}$). The prediction of $t_1$ is generated at the time of registration. (2) $t_2$: time from initial assessment to start-of-treatment ($t_2 = t_{treat} - t_{assess}$). The prediction of $t_2$ is generated at the time of initial assessment. Note that $t_1$ relates to an initial estimate of $t_{treat}$, and $t_2$ updates this estimate at the time of initial assessment, whereby information regarding triage and changes in patient-flow are incorporated in the modelling.

### 2.2 Patient-level ED data

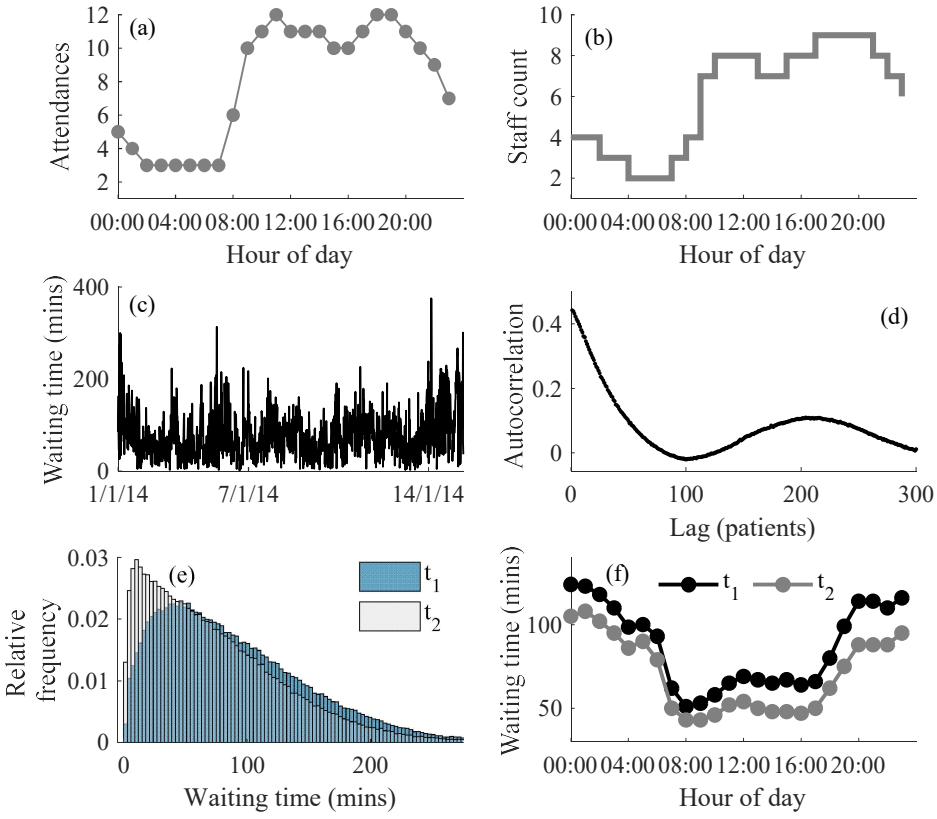
We employ five years of patient-level ED data from 1 January 2014 to 31 December 2018. The data from the JR is feature-rich. Recording multiple data fields for each patient while working in a high-pressure environment such as an ED, however, also increases the likelihood of having incomplete entries. It is rather unsurprising that EDs are particularly vulnerable to data quality issues (Ward et al. 2015). In practice, domain experts may employ several criteria to deal with data incompleteness. In this study, to avoid any potential bias resulting from data imputation, we adopt stringent criteria for pre-processing where we discard entries with either: incomplete or missing data fields (null entries), faulty timestamps (negative waiting times), or highly unlikely values (waiting times ≥14 hours, age ≥ 110 years). From a total of 454,983 patient-level records, we identified 352,178 patient-level records that were of good quality for the modelling. Out of the 352,178 patients visits: 35.8% were ambulance arrivals, 52.2% were female, 34.2% were admitted to the hospital, the average age was 39 years (standard deviation 27.2 years), 24% were minors (age ≤ 16 years), 25.7% were senior citizens (age ≥ 60 years). Low-acuity patients ($n$ = 334,635) constitute around 95% of the ED attendances, and are the focus of this study. Of these low-acuity patients, those triaged as minor and major injury comprised 43.6% and 51.4% of the total ED visits, respectively.

During the modelling, we first generate a probability distribution estimate of the waiting time from registration to the start-of-treatment ($t_1$) for each of the 334,635 low-acuity patients. We then update this prediction by estimating the waiting time from initial assessment to start-of-treatment ($t_2$) for the 281,910 patients that underwent an initial assessment. Note that, depending upon the severity of the condition and utilization of the ED, some patients (52,725 patients in our dataset) start their treatment

without first undergoing an initial assessment. The average start-of-treatment waiting time from point of registration ($t_1$) and from initial assessment ($t_2$) was 86.9 minutes (standard deviation 64.9 minutes) and 72.5 minutes (standard deviation 60 minutes), respectively. We use the first four years (2014–2017) as the in-sample period to estimate model parameters, while the final year (2018) was employed as the out-of-sample period to evaluate forecast accuracy.

Figure 2 presents median diurnal profiles of *attendances* (i.e., arrivals) and staff count, along with plots of the waiting times ($t_1$ and $t_2$). Attendances at EDs exhibit strong diurnal periodicity, usually with demand low during the night, and peaks around midmorning and early evening (Figure 2a). A similar diurnal pattern is observed for the median staff count, which is expected, as hospitals allocate more staff to the ED during busier periods of the day (Figure 2b). Figure 2c presents start-of-treatment waiting times ($t_1$) for a two-week period. It is evident that waiting times are highly variable, which underscores the need to generate probabilistic forecasts. Figure 2d presents the autocorrelation plot for waiting times ($t_1$), which is computed at different lags across consecutive patients. This figure shows that waiting times for consecutive patients are autocorrelated, which motivates our decision to employ empirical benchmarks that are based on the rolling average methods used by hospitals in practice.

**Figure 2.** Plots of attendances, staff count, and waiting times. Panel **2a**: median diurnal attendances, **2b**: median diurnal staff count, **2c**: time series plot for a fortnight of start-of-treatment waiting times from the time of registration ($t_1$), **2d**: autocorrelation function of $t_1$, **2e**: relative frequencies of start-of-treatment waiting times from the time of registration and time of initial assessment ($t_1$ and $t_2$, respectively), and **2f**: median diurnal waiting times ($t_1$ and $t_2$).



The waiting time distributions are right-skewed, as some patients end up waiting for many hours in the ED (Figure 2e). Waiting times are typically lower during hours of the day when more staff are available (Figure 2f). A plot of median diurnal waiting times for different days of the week is provided in the Supplement (Figure A1), which shows that waiting times are overall higher during weekends and Mondays, and tend to be lower on Thursdays. Figure 2 is generated using only the in-sample data.

## 3. Modelling Waiting Times Using a Quantile Regression Forest

In this section, we present our proposed modelling approach. Section 3.1 describes the process of feature engineering using the ED data. The features are used as input variables during the modelling. Section 3.2 presents the quantile regression forest method, which is an extension of random forests. Rankings of the most salient features derived from this approach are provided in Section 3.3.

### 3.1 Feature Engineering

Feature engineering is an integral part of modelling, as the accuracy of any statistical method or machine learning approach is conditional upon the quality of input features (Guyon et al. 2008). In the context of this study, feature engineering involves deriving potential predictors of waiting times from the raw patient-level ED data. The process of feature engineering can be either automated or manual. In this study, we employ manual feature engineering as it allows us to accommodate domain expertise and knowledge of the ED workload and patient symptoms during the modelling. Moreover, manually extracted features, which are amenable to interpretation, can potentially assist ED service providers to identify the main bottlenecks of waiting times. The patient-level records used in this study are detailed, which allows us to extract a range of features for each patient.

Table 1 presents a list of features extracted from the data along with a brief description. Each feature belongs to one of the following five categories:

(Category 1) **Calendar effects**: These features accommodate periodic variations in waiting times across the day (diurnal periodicity) and week (weekly periodicity). Waiting times are typically longer around 8pm-2am (Figure 2f), and during weekends and Mondays (Figure A1). Moreover, EDs often experience anomalous levels of attendances during the holiday periods (Rostami-Tabar and Ziel 2020). To incorporate the effect of anomalous load on waiting times, we use indicator variables to identify holidays (such as Christmas) and winter proximity days (days around Christmas day).

(Category 2) **Demographics**: These features account for potential differences in waiting times across different demographics (such as age, sex).

(Category 3) **Staff count**: This feature accounts for the ED service capacity due to staffing. Waiting times are typically low during periods of the day when more staff are available (Figure 2b and 2f). Due to data protection issues, information regarding staff schedules was not directly provided by the hospital, so we could not identify if a given staff member was either a nurse or a physician, or if they were permanent or temporary staff. Thus, following Ang et al. (2016), we employ the unique codes of the staff members responsible for discharging a patient to infer the total hourly staff count.

(Category 4) **ED operations**: These features reflect the state of the ED's operations at any given time, allowing us to represent changes in the ED workload at different points of patient-flow. Relevant features include the numbers of patients in the ED and in different status (e.g., registered but not yet assessed). See Table 1 for the full list of features. Each patient arriving in the ED is assigned a clinical triage category, which helps the ED to prioritize patients with time-critical injuries and better allocate resources. The JR broadly triages patients as: 'resus' (resuscitation), 'urgent care', 'major injury', and 'minor injury'. We refer to patients that need resuscitation or require urgent care as 'high-acuity' patients, as they are treated with priority upon their arrival at the ED. We refer to patients triaged as 'major injury' and 'minor injury' as low-acuity patients, and they are the focus of this study. For a given low-acuity patient, the waiting time depends on the number of high-acuity patients that are currently in the ED queue. It can be envisaged that a high-acuity patient would be prioritized over a low-acuity patient that has waited longer. Consideration of the total number of high-acuity patients during the process of feature engineering is thus imperative to allow for a more complete and accurate representation of the ED workload during the modelling. Thus, although we generate waiting time estimates for only the low-acuity patients, we use data for patients across all triage categories for feature engineering. Moreover, we include features that indicate the number of patients who breached the NHS four- and 12-hour waiting time targets (over the last 24 hours). We also use hourly averages of lagged waiting times as features.

(Category 5) **Patient condition**: These features accommodate the severity of the patient's condition (or presenting complaint) during the modelling. At initial assessment, the following three metrics are used to quantify the patient's condition: (1) Patient group number – indicates the reason for ED episode (e.g., road traffic accident) (2) Human resource group code – indicates the level of resources needed for the patient. (3) Triage category – patients are prioritized for treatment based on their triage level. We are not aware of any other study on predicting patient waiting times based on such detailed information on patient condition. This aspect of modelling leads to a key advantage of our approach – that forecasts can be refined over time as patient information is updated.

To forecast $t_1$, we employ features belonging to Categories (1)-(4), as this information is available at the time of registration. At the point of initial assessment, additional information regarding patient symptoms (Category 5 features) also becomes available. To forecast $t_2$, we thus incorporate features belonging to Categories (1)-(5). Also, ED workload can change considerably from the time a patient registers at the ED to the time they are assessed, especially during peak times. To incorporate the dynamic nature of patient flow in the ED, at the time of assessment, we recalculate the features for ED workload (Category 4 features) while modelling $t_2$. This makes our modelling framework of waiting times personalized, probabilistic, and dynamic. This will be discussed further in Section 6.1.

**Table 1.** Category, names, and a brief description of different features used for modelling.

| **Feature category and name** | **Brief description** |
|---|---|
| Category 1: **Calendar effects** | |
| Hour of day | Arrival hour of day (1 to 24) |
| Hour of week | Arrival hour of week (1 to 24×7) |
| Day of week | Arrival day of week (1 to 7) |
| Month of year | Arrival month of year (1 to 12) |
| Holiday period effects | Indicator variables to accommodate anomalous waiting times during holidays period (0: Normal day; 1: Holiday; 2: Days around Christmas day) |
| Category 2: **Demographics** | |
| Age | Patient age (0 to 110 years) |
| Sex | Patient sex (0: Male, 1: Female) |
| Category 3: **Staffing** | |
| Staff count | Total hourly staff count (inferred via unique staff codes) |
| Category 4: **ED Workload and Rolling Averages** | |
| Total workload | Number of patients in the ED (total, ambulance arrivals, other mode of arrival) |
| Workload from registration to initial assessment | Number of patients in the ED that have registered but have not been assessed (total, ambulance arrivals, other mode of arrival) |
| Workload from initial assessment to start of treatment | Number of patients in the ED that have been assessed but have not started treatment (total, ambulance arrivals, other mode of arrival, number of patients triaged as, minor, major, urgent care, and, resuscitation; and ambulance arrivals triaged as minor, major, urgent care, and, resuscitation) |
| Workload from start of treatment to departure | Number of patients in the ED that have started treatment but have not yet departed (total, ambulance arrivals, other mode of arrival, number of patients triaged as, minor, major, urgent care, and, resuscitation; and ambulance arrivals triaged as minor, major, urgent care, and, resuscitation) |
| 4-hour breach | Number of patients who waited > 4 hours (from registration to departure, over the last 24 hours) |
| 12-hour breach | Number of patients who waited > 12 hours (from registration to departure, over the last 24 hours) |
| Rolling average of waiting times | Average hourly lagged waiting times for same hour of the day (for 7 previous consecutive days) |
| Category 5: **Patient condition** | |
| Triage level | Category to determine patient's priority for treatment (minor, major, urgent care, and, resuscitation) |
| Human resource group codes | Code to reflect the level of resources needed by the patient (12 alphanumeric codes) |
| Patient group number | Code to identify the reason for ED episode (defined by the NHS as: road traffic accident, assault, deliberate self-harm, sports injury, firework injury, other accident, brought in dead, and other than above) |

### 3.2 Quantile Regression Forests

In this study, the aim of modelling is to estimate the probability distribution function $\hat{F}(y_i|X_i)$ for a target observation $y_i$ that is conditional on the corresponding feature vector $X_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$, where $y_i$ denotes the waiting time for the $i^{\text{th}}$ patient, and $X_i$ ($X_i \in \mathbb{R}^m$) is a $m$-dimensional vector of features that quantify properties of the patient, calendar effects, and state of the ED (i.e., features listed in Table 1). For a given patient, $y_i$ refers to $t_1$ at the time of registration, while at the time of initial assessment, $y_i$ refers to $t_2$. Given $n$ low-acuity patients having waiting times, $Y = \{y_1, y_2, \dots, y_n\}$ (the *label vector* with size $n \times 1$), and corresponding features represented by $\boldsymbol{X} = \{X_1, X_2, \dots, X_n\}$ (the *feature matrix* with size $n \times m$), we aim to train a model (denoted by $\Omega$, with parameter matrix $\boldsymbol{B}$), which is a mapping from the input features to the corresponding waiting time distribution, $\hat{F}(y_i|X_i) = \Omega(X_i, B_i)$. We provide a simplified and concise description of QRF proposed by Meinshausen (2006) so that the methodology is more accessible to a broader audience.

QRF is a generalization of the popular random forest (RF) method. RF is an ensemble machine learning method that has commonly been used to generate accurate predictions using high-dimensional features (Breiman 2001). The performance of RF has been shown to be robust under the presence of noisy or highly correlated features (Breiman 2001). Besides prediction, RF can be used to derive rankings of feature importance which can help make valuable inference from the data (Hastie 2009). While RF provides an accurate approximation of the conditional mean of the target variable in a nonlinear and nonparametric framework, QRF estimates the conditional probability distribution of the target variable.

To construct a QRF, we grow a large set of regression trees. While growing each tree and node, randomness is incorporated during the selection of features. For a given tree, a bagged version of the training data is used. For each node, a random subset of features is used for splitpoint selection while approximating the target variable. A tree is grown by splitting the bootstrap training sample such that it minimizes the total impurity (sum of squared deviations about a group mean). The process of splitting continues until a minimum leaf size has been achieved. Given the set of trees, dropping a new data point down each tree reaches a leaf node that produces a single forecast (observation) of the target variable. While RF estimates the conditional mean of the target variable by averaging such observations over the set of trees, QRF stores the value of all observations in the leaf nodes to estimate an empirical cumulative distribution function of the target variable.

For a single tree, denoted by say $T(\theta)$, which is grown using a random feature subset $\theta$, the point forecast of the mean obtained using a new feature vector $X_{new}$ is computed as the average of the subset of target values $y_i$ in the training sample associated with feature vectors $X_i \in l(X_{new}, \theta)$, where

$l(X_{new}, \theta)$ is the leaf node that contains $X_{new}$. Mathematically, the forecast of the mean is given by $\sum_{i=1}^{n} w_i(X_{new}, \theta)\, y_i$, where the weights $w_i(X_{new}, \theta)$ are given by:

$$w_i(X_{new}, \theta) = \frac{1_{\{X_i \in l(X_{new}, \theta)\}}}{\left|j: X_{j \in l(x,\theta)} \in l(X_{new}, \theta)\right|}.$$

For a forest of $n_{tree}$ regression trees, the weights from each tree are averaged as: $w_i(X_{new}) = 1/n_{tree}\left(\sum_{j=1}^{n_{tree}} w_i(X_{new}, \theta_j)\right)$, where $\theta_j$ denotes the feature subset used for growing the $j^{\text{th}}$ tree. In RF, the conditional mean $E(y|X = X_{new})$ can be estimated as $\sum_{i=1}^{n} w_i(X_{new})\, y_i$. QRF provides a natural extension for probabilistic forecasting by estimating the conditional distribution as:

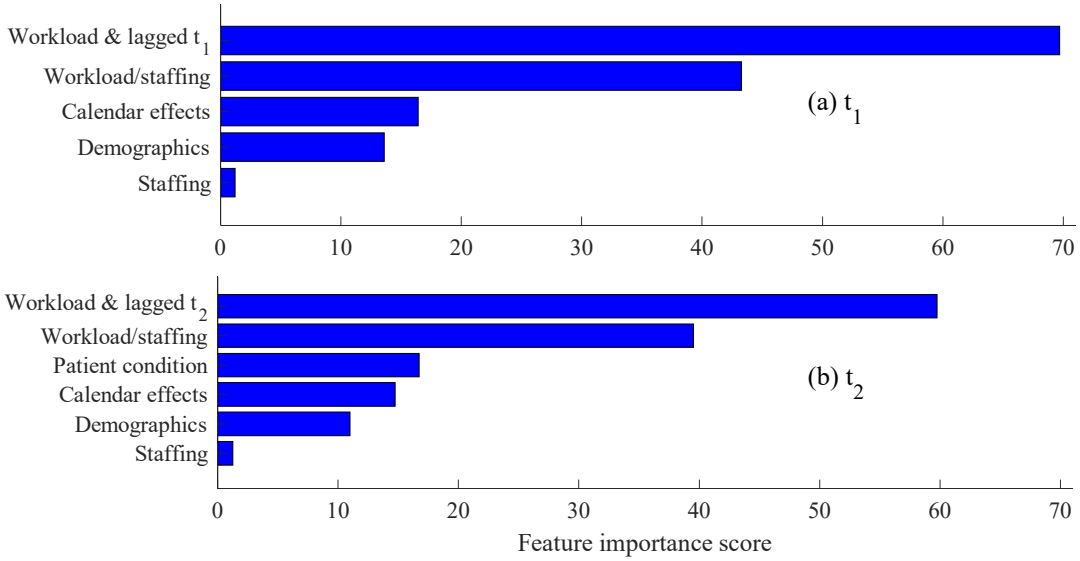
$$\hat{F}(y|X = X_{new}) = \sum_{i=1}^{n} w_i(X_{new}) I(y_i < y).$$

Instead of computing the average of observations in leaf nodes as done in RF, QRF keeps track of all observations and their weights, for all leaves and across all trees, to estimate the distribution function of the target variable. This estimate of the conditional distribution function is asymptotically consistent (Meinshausen 2006). In a recent study, using features extracted from flight- and passenger-level data, regression trees have been shown to be effective in predicting the probability distributions of connection times and the number of arrivals at an airport (Guo 2018). For QRF, we tuned three hyperparameters, namely the: (1) total number of trees in the ensemble, (2) minimum size of the leaf node for a given tree, and (3) minimum number of features used for split-point selection. The hyperparameters were tuned by minimising the mean absolute error (MAE) on the cross-validation hold-out sample. The last year of training data was used as the hold-out period. Using the hyperparameter values corresponding to the lowest MAE, we retrained the QRF using the whole in-sample period (2014-2017) and employed the trained model for generating forecasts on the out-of-sample period (2018). The QRF hyperparameter values are presented in the Supplement (Table A1).

### 3.3 Rankings of Feature Importance

We also use the RF underlying the QRF approach for identification of the most salient features associated with predicting waiting times at different stages of patient-flow in the ED. To calculate feature importance, the RF was re-estimated with one feature omitted, and the corresponding increase in error on the out-of-bag (OOB) observations was calculated. This was repeated for each feature. The most relevant features are judged to be those associated with the largest increases in the OOB error. The OOB errors were first calculated for each tree, and then averaged across all trees. Teasing out the set of features with the strongest predictive power can help service providers focus on a small set of factors that influence patient flow, and inform further studies to uncover nonlinear relationships.

**Figure 3.** Feature importance scores for the five feature categories, used for modelling: (a) $t_1$: registration to start-of-treatment, and (b) $t_2$: initial assessment to start-of-treatment.




*Note*: a higher score denotes a higher-ranked feature category. For a given feature category, we present the sum of importance scores for all features within that category. The above scores thus denote the total contribution of each feature category in modelling waiting times.

Figure 3 provides the feature importance scores for different categories of features (as presented in Table 1) obtained using the random forest of Section 3.2 applied to the in-sample data. For both $t_1$ and $t_2$, the most salient features are associated with ED workload due to patient volumes and rolling averages of waiting times, and workload/staffing count (i.e., features inspired from fluid model-based estimators as used in Ang et al. 2016). Interestingly, while the workload is one of the most important drivers of waiting times, staff count did not show up among the list of most salient features in Figure 3, which is broadly in agreement with the feature rankings reported by Ang et al. (2016). Ding et al. (2010) reported calendar effects (day of week and time of day) to be important predictors of waiting times, while Sun et al. (2012) found that after controlling for queue sizes and patient flow rates, calendar effects were not significant predictors of median waiting times. We found calendar effect features to be more important than features for patient demographics and staff count. Since both attendances and staff count exhibit a strong diurnal periodicity, it is possible that the diurnal variation in staff count is largely captured by features that quantify attendances (i.e., ED workload). These findings provide an insight into the importance of incorporating a detailed description of ED workload during the modelling of waiting times. A brief description of the salient features is provided in the Supplement (Tables A2 and A3).

## 4. Benchmark Methods

This section presents a range of naïve and sophisticated benchmarks for modelling waiting times. Specifically, as naïve benchmarks, we present three empirical methods in Section 4.1. As sophisticated benchmarks, we present quantile regression (Ding et al. 2010; Sun et al. 2012), Q-Lasso (Ang et al. 2016), $k$-nearest neighbour ($k$-NN), and boosted regression trees in Sections 4.2-4.5. Parameter estimates for the empirical benchmarks, Q-Lasso, and $k$-NN, are provided in the Supplement (Table A4). For parameter estimation, we used the last year of the in-sample data (2017) as the cross-validation hold-out sample. All models considered in this study were implemented using the Matlab software (R2019b; Mathworks®, USA).

### 4.1 Naïve Empirical Benchmarks

The empirical methods, although simplistic in their mathematical formulation, are associated with low computational complexity, which makes them attractive for deployment in the ED. We use the following three empirical methods (based on rolling averages) for forecasting waiting times:

(1) **Empirical 4-hour** – empirical distribution of waiting times observed in the previous four-hours. For EDs that publish waiting time estimates in the US, a four-hour rolling average has become the conventional choice (Dong et al. 2019).

(2) **Empirical $p$-hour** – waiting times for the last $p$ consecutive hours ($h$-1, $h$-2, …, $h$-$p$), where $h$ denotes the hour of arrival for the current patient.

(3) **Empirical $q$-period** – waiting times of the previous $q$ periods conditional on the same period of the day ($h$-24, $h$-2×24, …, $h$-$q$×24).

To produce these three estimates of the probability distribution of $t_1$, we use historical waiting times from the time of registration to start-of-treatment. For $t_2$, we employ waiting times from the time of initial assessment to start-of-treatment. We select $p$ and $q$ based on the minimum root mean squared error (RMSE) over the cross-validation hold-out sample. As opposed to QRF and our sophisticated benchmarks, the empirical benchmarks focus solely on the temporal correlation in waiting times.

### 4.2 Quantile Regression

Although least squares regression has been employed for forecasting waiting times (Asaro et al. 2007), it is limited because it only estimates the conditional mean. A Gaussian distribution could be assumed, but this is unappealing because waiting time distributions are heavily right-skewed (Sun et al. 2012), as evident from Figure 2e. Quantile regression has thus been proposed for modelling waiting times (Ding et al. 2010; Sun et al. 2012), as it allows for a more detailed characterization of the ED data

by quantifying the impact of features on the distribution of waiting times. In this study, we estimate quantile regression models for the following quantiles τ = {5%, 15%, 25%, 35%, 45%, 50%, 55%, 65%, 75%, 85%, 95%}. To construct distributional forecasts using quantile regression, we linearly interpolate between the estimated quantiles and treat the minimum and maximum of the in-sample data as bounds of the distribution. The method could be considered rather cumbersome to implement, with the need for a model for each prespecified quantile, and an interpolation method.

### 4.3 Q-Least Absolute Shrinkage and Selection Operator (Q-Lasso)

This method was proposed by (Ang et al. 2016) for generating point estimates of patient waiting times. Q-Lasso combines concepts from queuing theory with the statistical modelling approach of Lasso, which employs the following objective for model estimation:

$$min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta^T X_i)^2 + \lambda ||\beta||_1$$

where $||\beta||_1 = \sum_{j=1}^{m} |\beta_j|$ is the L1 norm of the coefficients $\beta$ ($\beta \in \mathbb{R}^m$), and $\lambda$ is the regularization parameter. The penalty term $\lambda||\beta||_1$ helps prevent model over-fitting by forcing coefficients of less salient features to go to zero. Lasso thus makes the regression model more parsimonious. For $\lambda$ equal to 0, Lasso defaults to ordinary least squares. This method predicts waiting times as a linear function of features within a parametric modelling framework. We estimate $\lambda$ using 5-fold cross-validation (for details, see Hastie 2009). We used this method to generate only point forecasts as there are challenges associated with Lasso when estimating standard errors of regression coefficients (Goeman 2010). It is worth noting that Lasso assumes a linear relationship between the features and the waiting times. However, given the complex and dynamic nature of patient-flow in an ED, it can be envisaged that such an assumption is too restrictive.

The Q-Lasso method proposed by Ang et al. (2016) incorporates fluid model estimators as candidate predictors in Lasso. In the context of modelling patient waiting times, these estimators generalize to the ratio: *workload/processing rate*. Workload is associated with the number of patients that must be seen before a new low-acuity patient can start treatment (Category 4 features), whereas the processing rate depends on the number of available staff (Category 3 feature). Following Ang et al. (2016), for a given workload feature (say, total workload), an additional feature was included in the model, which was calculated by dividing that feature with the corresponding hourly staff count (e.g., total workload/staff count). These features, which were calculated by dividing workload features by staff count, were used as inputs for all models considered in this study.

### 4.4 $k$-Nearest Neighbour ($k$-NN)

Although it has not been used before in the literature for forecasting waiting times, we use $k$-NN as it can be interpreted as a more sophisticated adaptation of the rolling average methods that are used in practice. While rolling average methods rely on averaging the observations of waiting times (for say, the last $p$-hours, or $q$-periods), $k$-NN generates waiting time estimates for a given patient based on the waiting times of the previous $k$ most similar patients. Similarity between any two patients is quantified by Euclidean distance in the feature space. As the five categories of features constitute a different number of features (Table 1), Euclidean distance was computed separately for each feature category, to ensure that each received equal weight during the modelling. Features were standardized to have zero mean and unit standard deviation. For each patient, probabilistic forecasts were generated by using the empirical distribution of the waiting times for the historical $k$ most similar patients. We estimate $k$ by minimizing the RMSE over the cross-validation hold-out-sample. While $k$-NN allows a nonlinear and nonparametric modelling of the waiting times, it does not provide insight into the rankings of most important predictor variables.

### 4.5 Boosted Trees

We also employed boosted trees as a sophisticated benchmark. We are not aware of their previous use for waiting time modelling in the literature. In the context of this study, the rationale of boosting is to build iteratively an ensemble of weak regression trees, such that waiting times that are more challenging to predict are assigned a higher weight. Specifically, using this methodology, waiting times associated with a large prediction error at a given iteration are assigned a higher weight for the next iteration. By sequentially changing the weights, the trees are made to concentrate on those patients for whom the waiting times are harder to predict. Predictions from individual trees are combined using a weighting strategy to issue a final forecast. Due to its relative ease of implementation, we use least squares boosting (LSBoost) that has been used previously for point forecasting (Friedman 2001). Compared to QRF, the hyperparameter tuning process for LSBoost comprises one additional parameter, namely the learning rate. As with QRF, the hyperparameters for LSBoost were tuned using the cross-validation hold-out sample. While boosted trees are attractive for forecasting tasks, they are typically harder to tune than RF because boosting builds the trees sequentially and requires the estimation of an additional parameter (learning rate), which can make boosting more prone to over-fitting. Moreover, for categorical forecasting of waiting times, we use random undersampling boosting (RUSBoost) as it has been shown to be effective in dealing with imbalanced datasets (see, for example, Seiffert et al. 2010). Probabilistic forecasting using gradient boosting typically requires building multiple models for

each separate quantile, or calculating second-order derivative statistics (Duan et al., 2020), which can be computationally expensive especially for high dimensional feature matrices and large datasets. We thus use boosted trees for point and categorical forecasting, and not for forecasting the full distribution.

## 5. Evaluating Forecasts

In this section, we compare the performances of QRF and the benchmark methods in terms of their point, quantile, categorical, and probability distributional forecast accuracy. Model rankings are assessed based on the accuracy of forecasting $t_1$ and $t_2$, using data for the one-year out-of-sample period (2018). Using our more detailed ED dataset, we were able to extract a richer set of features compared to previous studies. For example, to forecast patient waiting times, Sun et al. (2012) included information regarding the date, time of triage, time of consultation, and patient acuity category as features for a quantile regression model, and Ding et al. (2010) employed only two crowdedness measures: the ED occupancy rate and the bed occupancy rate. In their use of Q-Lasso, Ang et al. (2016) used only patient triage category to quantify patient condition. To ensure a fair evaluation of different forecasting strategies, we employ all the features of Table 1 in all methods, with the exception of the naïve benchmarks. Forecasts for $t_1$ were generated using features from the first four categories. For forecasting $t_2$, features from all five categories were used for the modelling.

### 5.1 Evaluating Point Forecasting

In Figure 4, we plot the median of the out-of-sample distributional forecasts of QRF (for $t_1$ and $t_2$) across different hours of the day using the one-year out-of-sample period. This figure shows that QRF is overall able to accommodate the diurnal periodicity in waiting times. For a quadratic loss function, the mean is the optimal point forecast, whereas the median is optimal if the loss function is symmetric piecewise linear (Gneiting 2011). In view of this, we use the RMSE and MAE to evaluate point forecasts obtained as the mean and median, respectively, of each distribution. For Q-Lasso, only one form of point forecast was produced, so we evaluated this using both RMSE and MAE.

**Figure 4.** Median diurnal variations in out-of-sample actual and predicted waiting times for: (a) $t_1$: registration to start-of-treatment, and (b) $t_2$: initial assessment to start-of-treatment.

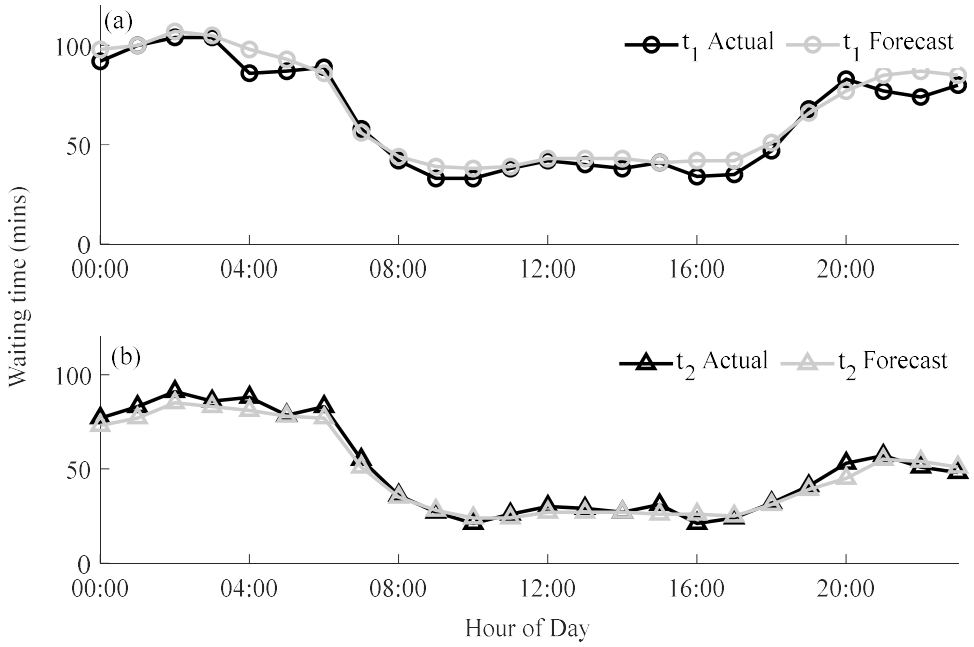



**Table 2.** RMSE and MAE for point forecasting of: (a) $t_1$: registration to start-of-treatment, and (b) $t_2$: initial assessment to start-of-treatment.

| | **RMSE** | | **MAE** | |
|---|---|---|---|---|
| **Forecasting Method** | $t_1$ | $t_2$ | $t_1$ | $t_2$ |
| Empirical 4-hour | 60.4 | 57.5 | 42.5 | 40.2 |
| Empirical $p$-hour | 55.9 | 53.3 | 39.8 | 37.2 |
| Empirical $q$-period | 63.4 | 62.3 | 43.9 | 44.3 |
| Quantile regression | 50.8 | 49.0 | 36.0 | 33.9 |
| Q-Lasso | 51.0 | 48.6 | 37.1 | 33.0 |
| $k$-nearest neighbour | 60.7 | 57.4 | 45.1 | 41.9 |
| LSBoost | 51.0 | 47.3 | 36.7 | 32.8 |
| QRF | **49.9** | **46.5** | **34.9** | **31.5** |

*Note*: lower RMSE and MAE values are better (lowest values are highlighted in **bold**).

In Table 2, we present the RMSE and MAE results. For point forecasting, QRF is the best performing method. Interestingly, LSBoost outperforms the sophisticated benchmarks, Q-Lasso and quantile regression, which were proposed in the literature for modelling waiting times. The performances of Q-Lasso and quantile regression are quite similar. The poorest results were produced by the empirical benchmarks and $k$-NN.

In addition to the RMSE and MAE, we evaluate out-of-sample point forecasts using a practical performance measure, which calculates the percentage of patients for whom the waiting times were

correctly estimated within 20, 40, and 60 minutes of the actual waiting times. The results are presented in Table 3. The performance of QRF was the best in estimating the waiting times within 40 and 60 minutes. For estimation within 20 minutes, the best performing methods were the empirical q-period and Q-Lasso methods for t_1 and t_2, respectively. We note, however, that, along with the RMSE and MAE, this performance measure evaluates only point forecast accuracy. In the remainder of Section 5, we consider the evaluation of probabilistic forecast accuracy.

**Table 3.** Percentage (%) of patients in the out-of-sample period for whom the predicted waiting times are correctly estimated within 20, 40 and 60 minutes of the actual times for: (a) $t_1$: registration to start-of-treatment, and (b) $t_2$: initial assessment to start-of-treatment.

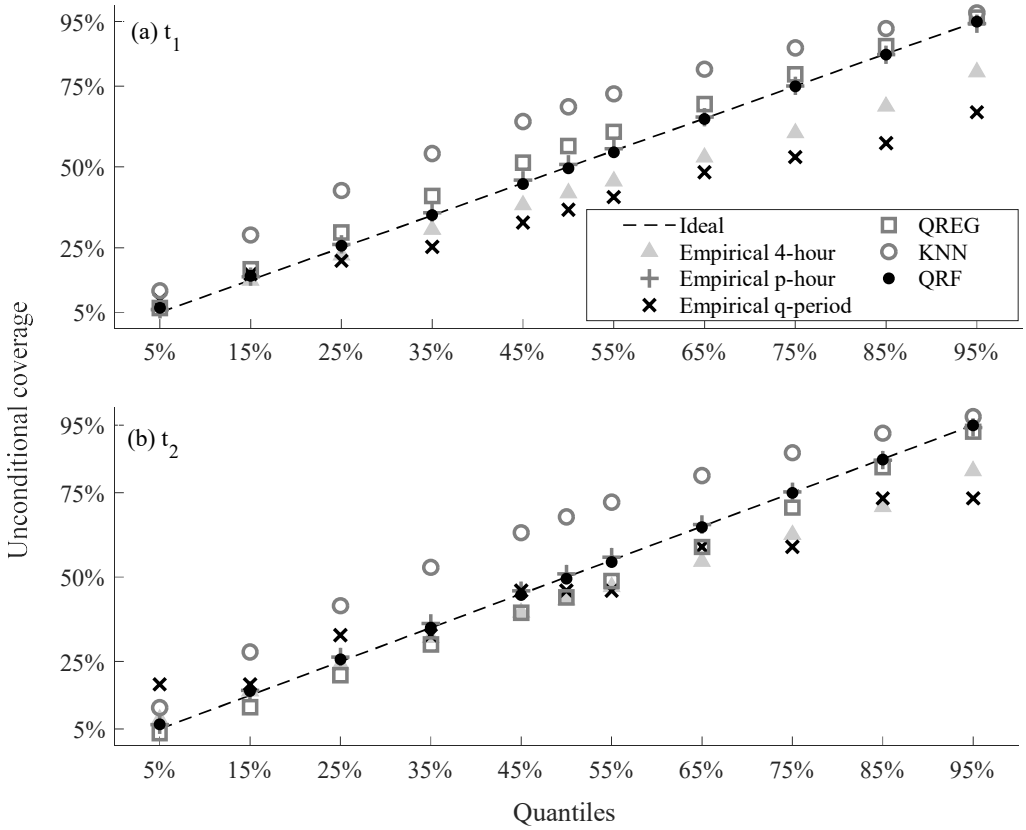
| | $\lvert t_1 - \hat{t}_1 \rvert$ | | | $\lvert t_2 - \hat{t}_2 \rvert$ | | |
|---|---|---|---|---|---|---|
| | **< 20 mins** | **< 40 mins** | **< 60 mins** | **< 20 mins** | **< 40 mins** | **< 60 mins** |
| Empirical 4-hour | 37.3% | 63.7% | 78.2% | 39.5% | 66.6% | 80.3% |
| Empirical $p$-hour | 32.4% | 62.6% | 78.7% | 33.8% | 65.9% | 81.6% |
| Empirical $q$-period | **39.8%** | 64.3% | 77.3% | 37.7% | 62.2% | 76.3% |
| Quantile regression | 39.6% | 67.5% | 82.6% | 44.7% | 70.5% | 84.1% |
| Q-Lasso | 35.7% | 65.5% | 82.4% | **44.8%** | 72.0% | 85.2% |
| $k$-nearest neighbour | 22.1% | 46.3% | 69.5% | 23.1% | 49.2% | 73.2% |
| LSBoost | 36.6% | 66.8% | 82.8% | 43.3% | 72.4% | 86.0% |
| QRF | 36.9% | **67.6%** | **83.4%** | 43.1% | **73.4%** | **86.7%** |

*Note*: higher percentage (%) values are better (highest values in each column are highlighted in **bold**).

### 5.2 Evaluating Quantile Forecasts

To evaluate quantile forecasts, we use *unconditional coverage*, which measures the percentage of observations that are lower than the $\tau$ quantile forecast. Ideally, this percentage should be $\tau$. Figure 5 presents the unconditional coverage for $t_1$ and $t_2$, averaged across all low-acuity patients, for $\tau$ = 5%, 15%, 25%, 35%, 45%, 50%, 55%, 65%, 75%, 85% and 95%. In Figure 5, values closer to the diagonal line (ideal coverage) are better. It can be seen from the figure that the unconditional coverage is rather impressive for empirical $p$-hour, quantile regression (QREG), and QRF. The performances are relatively poor for the other two empirical benchmarks and $k$-NN.

**Figure 5.** Unconditional coverage for quantile forecasts of waiting times for: (a) $t_1$: registration to start-of-treatment, and (b) $t_2$: initial assessment to start-of-treatment. *Note*: values closer to the diagonal are better.




### 5.3 Evaluating Probability Distributional Forecasts

Figure 6 provides a representation of distributional forecasts generated by QRF for two days chosen at random from the out-of-sample period. The forecast origin was midnight at the start of 8 April 2018 (Sunday). The median of the forecast distribution was issued as the point forecast. The point forecast and its uncertainty exhibit a similar diurnal pattern to those shown in Figure 2f. Figure 6 shows the 90% interval of the forecast distribution encompassing the vast majority of the actual observations.

To assess distributional forecast accuracy, we use the continuous ranked probability score (CRPS), which quantifies both *sharpness* (concentration or peakedness of the forecast distribution) and *calibration* (statistical consistency between forecast distribution and actual observations). We use the expectations form of the CRPS (Gneiting and Raftery 2007):

$$CRPS = E_F|Y - y| - \frac{1}{2} E_F|Y - Y'|$$

where $Y$ and $Y'$ are independent sampled values drawn from the forecast probability distribution function $F$ (we draw 1000 values), $E_F$ is the expectation with respect to the distribution $F$, and $y$ is the actual waiting time. An attractive property of the CRPS is that it is a strictly proper scoring rule, which means that it is minimized by the true distribution.

**Figure 6.** Summaries of out-of-sample QRF distributional forecasts for: (a) $t_1$: registration to start-of-treatment, and (b) $t_2$: initial assessment to start-of-treatment. *Note*: the shaded regions correspond to 90% and 50% intervals obtained from the forecast distribution.

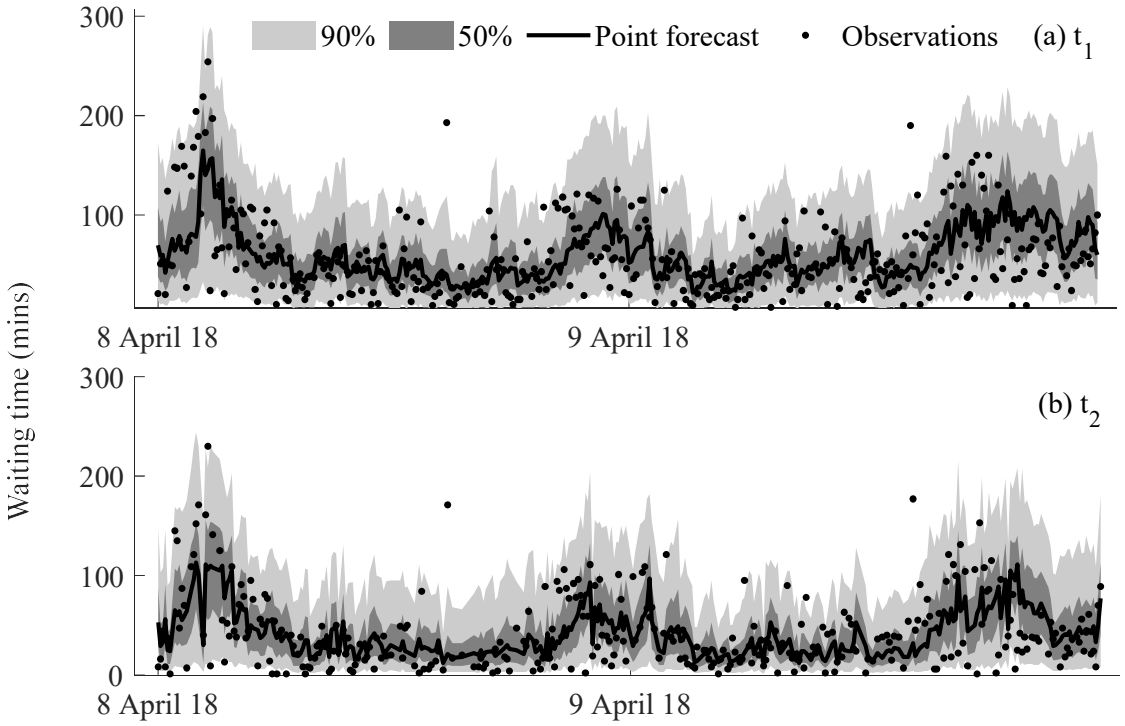



Table 4 presents the CRPS averaged across all patients in the out-of-sample period. QRF is the most accurate method for forecasting both $t_1$ and $t_2$. QRF provides a reduction of more than 20% in the CRPS compared to the empirical 4-hour (typically used in practice). QRF also outperformed the other benchmark methods, though the improvement over quantile regression was relatively modest. Surprisingly, $k$-NN was outperformed by the empirical $p$-hours method. As explained earlier, we were not able to produce probabilistic forecasts from Q-Lasso. In Table 4, the CRPS values for $t_2$ are lower than the corresponding values for $t_1$. This implies greater accuracy as one gets closer to the start of treatment. This is because: (1) for any given patient, $t_2$ is less than $t_1$ by construction, (2) features that quantify ED workload due to patient workflow are updated at the time of assessment, and (3) crucially, additional features (that quantify the severity of the patient's condition) are incorporated while modelling $t_2$.

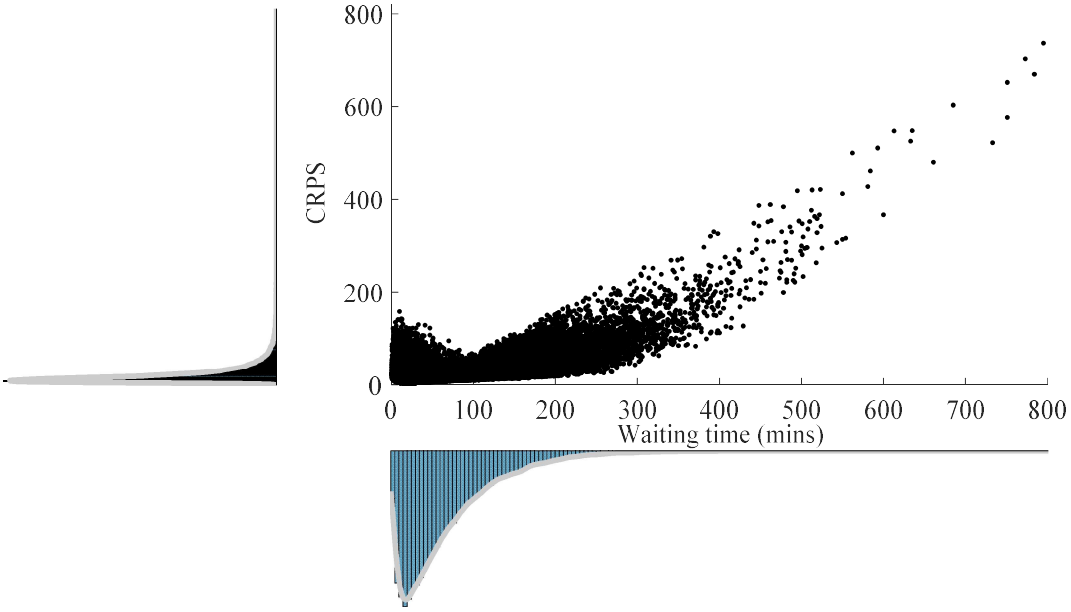
Figure 7 presents the out-of-sample CRPS values for each patient plotted against their corresponding actual waiting times. The magnitude of probabilistic forecast error is notably larger for patients that wait for exceedingly long hours in the ED. The accuracy is best for patients for whom the waiting time was close to the median in-sample waiting time of 73 minutes.

**Table 4.** Mean CRPS (and 95% CI of the mean CRPS, in parentheses) for distribution forecasting of: (a) $t_1$: registration to start-of-treatment, and (b) $t_2$: initial assessment to start-of-treatment.

| **Forecasting Method** | $t_1$ | $t_2$ |
|---|---|---|
| Empirical 4-hour | 31.7 (31.4-32.0) | 29.8 (29.4-30.1) |
| Empirical $p$-hour | 28.0 (27.7-28.2) | 26.3 (26.0-26.5) |
| Empirical $q$-period | 34.4 (34.1-34.7) | 34.2 (33.8-34.5) |
| Quantile regression | 25.5 (25.3-25.7) | 24.3 (24.0-24.5) |
| $k$-nearest neighbour | 31.3 (31.1-31.5) | 29.2 (28.9-29.4) |
| QRF | **24.7** (24.5-24.9) | **22.4** (22.2-22.7) |

*Note*: lower CRPS values are better (lowest values are highlighted in **bold**). CI: confidence interval.

**Figure 7.** CRPS values for all patients in the out-of-sample period plotted against their corresponding actual waiting time ($t_1$, in minutes). The left and bottom panel figures present the probability distributions of the CRPS and waiting times, respectively.




To gauge the effect of hyperparameter tuning, we compared the forecasting performance of two QRF models, with: (1) tuned hyperparameter values (based on minimization of the MAE on the cross-validation hold-out sample), and, (2) default hyperparameter values without tuning, which are 500 regression trees (Breiman 2001), one-third of the total features used for split-point selection (Meinshausen 2006), and nodes with less than 5 observations are not split any further. The out-of-sample performance of the tuned QRF was slightly better than the default QRF. Overall, QRF forecasting performance was robust to slight variations in the hyperparameters, as shown in the Supplement (Table A5).

### 5.4 Evaluating Categorical Probability Forecasts

As interpreting probability distributions is not straightforward for patients without statistical knowledge, we propose the following colour-coded reporting scheme: *Green* (for *low* waiting times $\leq$ 45 minutes), *Amber* (for *medium* waiting times: 45 minutes $<$ waiting times $\leq$ 120 minutes), and *Red* (for *high* waiting times $>$ 120 minutes). Converting the continuous target variable (waiting times) into a categorical variable (green, amber, red) transforms the modelling challenge from a regression problem into a classification task. Specifically, the aim of this task is to predict the probability of each of the three categories of waiting times. These probability forecasts can be obtained from the forecasts of the continuous probability distributions. We do this using the probability distributions of the different methods (compared in Section 5.3). We also considered three classifiers: RUSBoost, a three-class classifier, and a multiple binary classifier. For multi-class and binary classification, we use a random forecast classifier. Converting the waiting times into a categorical variable resulted in imbalanced data, i.e., a different number of training observations belonging to the three classes. Data imbalance makes the learner more prone to over-classify the majority class. To tackle this issue, we assign a larger weight to the underrepresented class (see, for example, He and Garcia 2009). To evaluate categorical forecasts, we use the ranked probability score (RPS), which is the score for discrete distributions that is analogous to the CRPS (Epstein 1969). Table 5 presents the average out-of-sample RPS. Encouragingly, the table shows QRF outperforming other modelling schemes in the classification task, including RUSBoost and the other two classifiers.

**Table 5.** RPS (×100) for categorical forecasting of: (a) $t_1$: registration to start-of-treatment, and (b) $t_2$: initial assessment to start-of-treatment.

| **Forecasting Method** | $t_1$ | $t_2$ |
|---|---|---|
| RUSBoost classifier | 16.1 | 14.9 |
| Three-class classifier | 15.6 | 14.0 |
| Multiple binary classifiers | 15.6 | 14.0 |
| Empirical 4-hour | 19.5 | 18.5 |
| Empirical $p$-hour | 17.2 | 16.2 |
| Empirical $q$-period | 21.5 | 21.1 |
| Quantile regression | 15.7 | 14.4 |
| $k$-NN | 19.6 | 18.3 |
| QRF | **15.2** | **13.6** |

*Note*: Lower RPS values are better (lowest values are highlighted in **bold**). The first three methods in the above table generate categorical probability forecasts, while the remaining methods produce continuous probability distribution forecasts, which we convert into categorical forecasts.

## 6. Implementation

In this section, we discuss aspects related to the practical implementation of the proposed probabilistic modelling framework. Section 6.1 focusses on publishing and updating waiting time estimates at different stages of patient-flow in the ED. Section 6.2 demonstrates that estimates of waiting times, when used in conjunction with travel times, could help patients make informed decisions while selecting an ED site from a network of neighbouring EDs.

### 6.1 Implementing the Forecasting Scheme at the ED with Information Updates

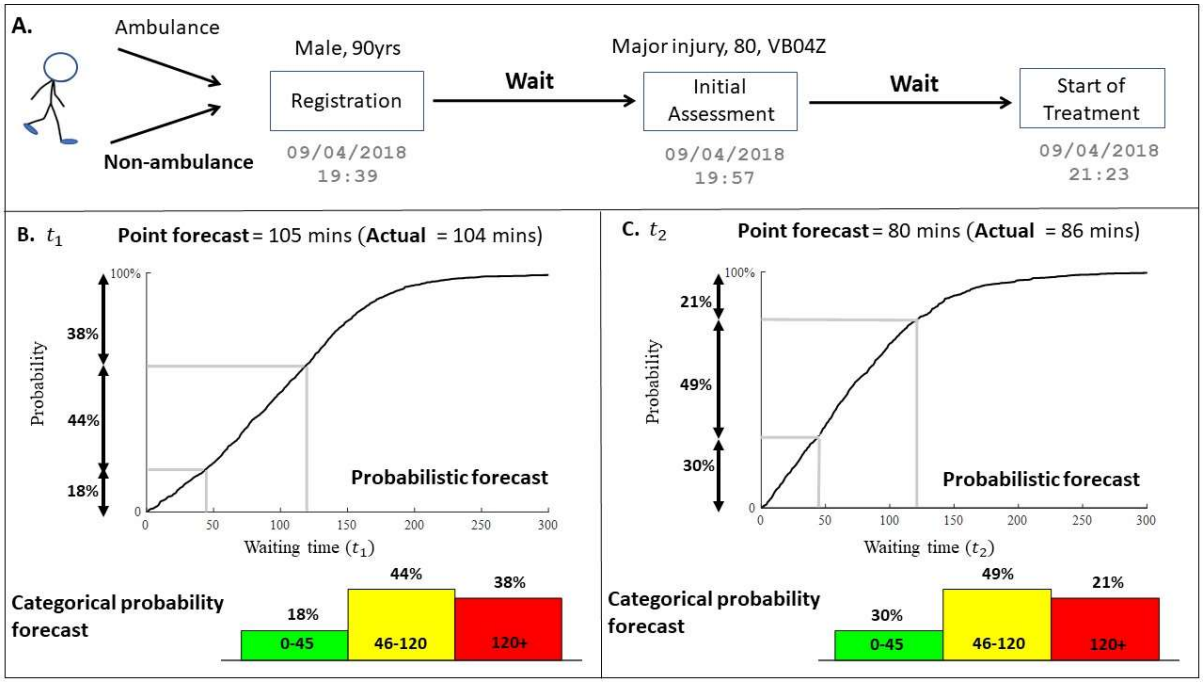
We now show how the proposed forecasting scheme could be used to communicate personalized waiting times, which patients could check using, say, a smartphone app/website or a TV screen in the ED waiting room. For demonstration, we select data for an actual patient from the out-of-sample period, and for ease of explanation, we refer to this patient as John.

Figure 8 demonstrates the patient-flow for John, who registered at 19:39 on 9 April 2018 (non-ambulance arrival). Using features for calendar effects (time and day of arrival etc.), demographics (90-year-old male), staff count, and state of the ED patient workload at John's time of registration (number of total patients, ambulance arrivals, etc), we generate the forecasts for $t_1$, the waiting time until treatment. The QRF point forecast indicated that John would start treatment in 110 minutes from the time of registration (actual $t_1$ was 104 minutes). As shown, QRF categorical forecasts were derived from the forecast of the continuous probability distribution. The colour-coded scheme for QRF categorical forecasts conveyed an 18% chance of waiting less than 45 minutes, and a likelihood of 44% for waiting between 46 to 120 minutes, and a 38% probability of waiting longer than 120 minutes. Note that the estimate for $t_1$ was generated at the time of registration.

At 19:57, John underwent initial assessment, and at this point he was triaged as 'major injury' and assigned a patient group number '80' (code denoting reason for presenting complaint) and a code 'VB04Z' (code denoting use of resources). This additional triage related information was incorporated into the model for $t_2$, the waiting time between initial assessment and treatment. Moreover, since the state of patient-flow at the ED can change rather quickly, we update feature values for the ED workload (Category 4 features) to accommodate changes in the ED from 19:39 to 19:57. The time spent from registration to the start-of-treatment was also included as a feature in the model for $t_2$. The updated point forecast from QRF indicated that John would start treatment in 79 minutes from the time of initial assessment (actual $t_2$ was 86 minutes). The updated colour-coded scheme shows a 30% probability of starting treatment within the next 45 minutes, a 49% probability of starting treatment between 46 to 120

minutes, and only a 21% probability of starting treatment after two hours. Note the initial forecast was generated at the time of registration, and updated at the time of initial assessment.

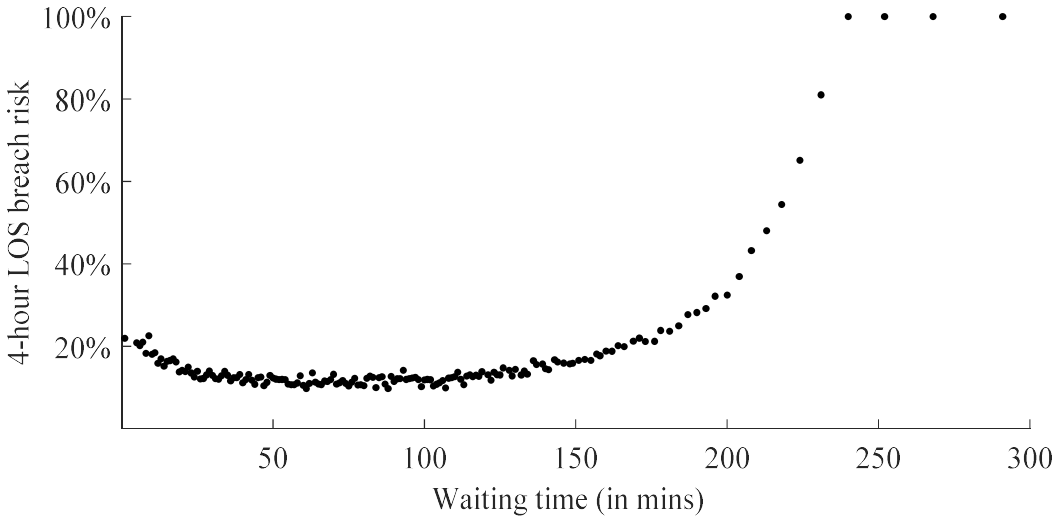
**Figure 8.** Schematic diagram illustrating ED patient-flow and forecasts. Panel **A** presents the actual timestamps of registration, initial assessment, and treatment (for a 90-year-old male, triaged as 'major injury', patient group '80' and code 'VB04Z' during the initial assessment). Panel **B** and Panel **C** shows the point forecast, probabilistic forecast, and categorical probability forecast, for $t_1$ and $t_2$, respectively. Forecasts shown in Panel B and Panel C are generated at the time of registration ($t_{reg}$) and time of initial assessment ($t_{assess}$), respectively.




The proposed methodology can be used to generate forecasts for patients in real-time. Using a trained QRF, the CPU time to generate waiting time estimates for a new patient at the time of registration was around 3 seconds, which was calculated as the average CPU time for 100 randomly chosen patients (using MATLAB R2019b; Mathworks®, USA, on an i7 processor with 32 GB RAM).

While estimates of waiting times are of particular interest to patients, EDs are generally more concerned with the total length of stay (LOS), which is defined as the time elapsed from a patient's registration to departure from the ED. In England, EDs are given a target of 4-hour LOS, within which at least 95% of the patients attending the ED should be treated, admitted or discharged (NHS 2019). Inability to meet this target can have financial and reputational repercussions for the hospitals. In this study, using the start-to-treatment waiting times ($t_1$), we devise a 4-hour LOS breach risk score, as shown in Figure 9 (generated using only the training data). As expected, a higher start-to-treatment waiting time is associated with a greater risk of the four-hour LOS target being breached.

**Figure 9**. Risk (%) of breaching the 4-hour length of stay (LOS) as a function of waiting time ($t_1$).




Using Figure 9, EDs could convert forecasts of start-to-treatment waiting times into the risk of the 4-hour target LOS being breached. For example, the point forecast of $t_1$ was 105 minutes for John (as shown in Figure 8), the corresponding risk of a 4-hour LOS breach at the time of registration was calculated as 10.8% (using Figure 9). Although beyond the scope of this study, an interesting line of future work would be to compare modelling LOS directly using features that characterize ED patient-flow versus modelling LOS as a function of $t_1$.

### 6.2 Implementing the Forecasting Scheme Remotely: A Demonstration

Studies have emphasized the importance of informed routing decisions to get the right patient to the right provider at the right time (Singh et al. 2020). Access to waiting times for different healthcare providers can assist patients and first responders to make better decisions in selecting a hospital site, which in turn has been shown to reduce actual waiting times (Xie and Youash 2011). Forecasts of waiting times, when used in conjunction with travel times, could potentially be used for ambulance routing and diversion, to facilitate the uniform spread of load within a network of neighbouring hospitals (Deo and Gurvich 2011; Xu and Chan 2016). To assist patients in selecting an ED from a network of EDs with different waiting times, smartphone applications have been proposed (such as, Waitless, NHS Quicker). However, the aforementioned studies and smartphone applications do not take into account the uncertainty in travel and waiting times.

For patients arriving by ambulance to the ED, the routing decision is typically made by emergency medical technicians (EMTs). As pointed out by Akşin et al. (2021), while taking a patient to the ED, ambulance crew in England select the geographically closest ED, and subsequently follow the shortest driving route to it. For a comprehensive survey on the ambulance routing problem, see Tassone and

Choudhury (2000). It is worth noting that in current practice in the NHS, personalized waiting times are not forecasted and thus this information is not accessible to either EMTs or patients or multiple ED sites within a geographic network. We show that decision-makers, such as patients and EMTs, could select an ED from a network of EDs based on probabilistic modelling of travel and waiting times. This analysis necessitates data for more than one ED. We include additional data from the Horton General (HG) hospital (situated in Banbury, UK, about 20 miles from the JR). HG and JR are the only two ED sites in Oxfordshire, a region with about 700,000 residents. Specifically, we use five years of data from 1 January 2014 to 31 December 2018 from the HG, comprising 180,715 patient-level records.

As with our study of the JR hospital data, the HG data for the first four years (2014–2017) was used for training, while data from the final year (2018) was used for evaluation. Encouragingly, the out-of-sample model comparison using data from the HG hospital revealed that QRF generated the most accurate forecasts of the benchmark methods that we considered in Section 4, consistent with our findings for the JR. The out-of-sample forecast errors for the HG hospital were around 20% less compared to the JR (Supplement Table A6), whereby the MAE for HG was around 25 minutes.

Previous studies have typically relied on building a separate model for each ED site (Ang et al. 2016). The advantages of harmonising and integrating healthcare datasets from multiple sources have, however, been highlighted (Detmer et al. 2008). In the context of this study, a centralized routing decision-support system built using an integrated dataset from multiple ED sites is attractive, as it could potentially facilitate interoperability and cooperation between EDs, a better understanding of patient flow, and efficient data management. Moreover, an integrated dataset could provide insight into temporal variations in patient flow across multiple ED sites in near real-time, which could assist patients and EMTs make informed and dynamic routing decisions. In this study, we investigate the efficacy of integrating ED datasets from the JR and the HG. Specifically, we train a single QRF model by integrating patient-level records from the JR and the HG. During the modelling, we include a binary variable as a feature to indicate the ED site. Using the integrated model, we generate out-of-sample predictions and evaluate the model performance separately for each ED site. We were interested to find that the forecast performance of the two modelling approaches: (1) using a single model constructed using an integrated dataset, and (2) a separate model built for each ED site; was rather similar (as presented in the Supplement, Table A6).

Using the QRF trained with the integrated dataset from the JR and the HG, we simulate both patient and ambulance/EMT choices while selecting an ED site by accommodating uncertainty in travel and waiting time estimates. Due to data protection issues, we did not have access to patient locations. Thus, for demonstration purposes, we randomly assign patients to postcodes from the geographic

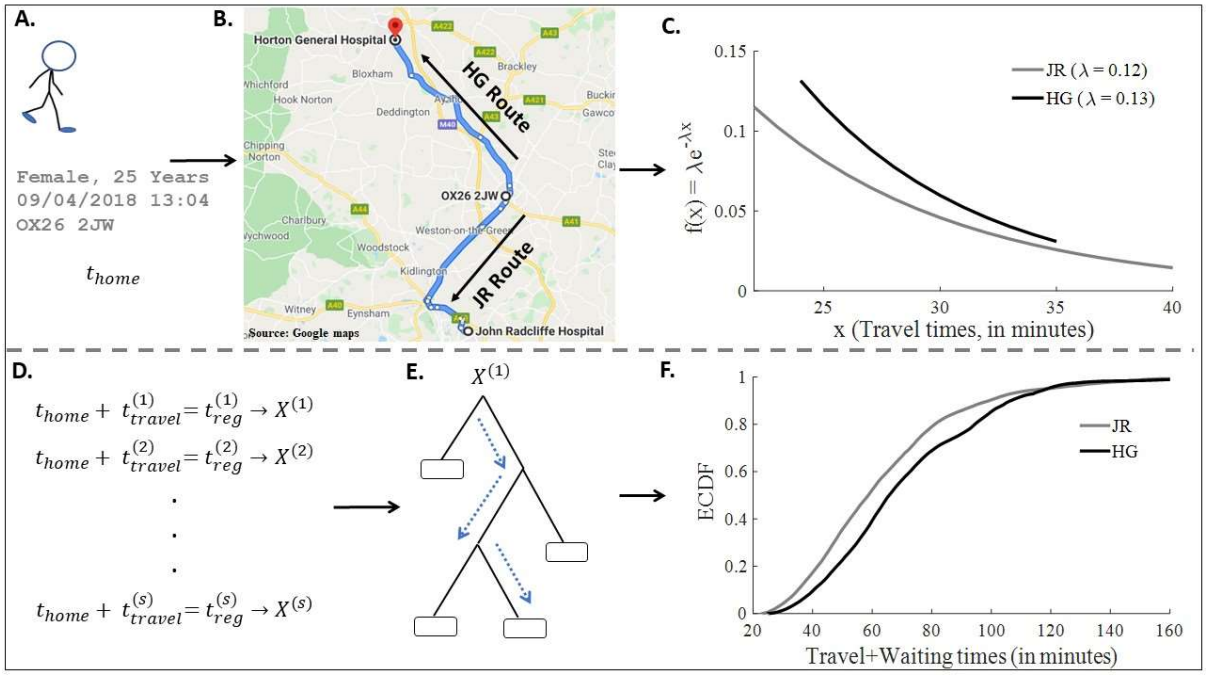
neighbourhood of the hospital sites. To estimate travel times, using each patient's timestamps (time of selecting an ED) and postcode, we accessed the following data from Google maps for both hospitals: distance, minimum driving time, and maximum driving time. Since Google maps do not provide driving estimates for ambulances, we used travel times for the previous midnight as a proxy for ambulance driving times. For simplicity, we assumed travel times are exponentially distribution. The minimum travel time was treated as the true lower bound, and we set maximum travel time as the 99% quantile of the exponential distribution. We simulate data for a day (9 April 2018) in the out-of-sample period for both hospitals ($n$ = 318 patients, which corresponds to the actual total attendance for the JR and HG for the day under consideration). Patients and EMTs could remotely access waiting time and driving time estimates using, say, a smartphone to make a routing decision to an ED site.

Figure 10 presents a schematic diagram that illustrates patient decision-making while selecting an ED site from a network with two hospitals. For example, consider a patient, say Anna, who needs to decide whether she should visit the JR or the HG. For her location (postcode: OX26 2JW), and time of decision while at home (13:04 on 9 April 2018, say $t_{home}$), we access traffic information regarding the travel time to both hospitals. This information is used to estimate the probability distribution of travel times (Figure 10C). We draw a random sample of travel times (denoted by $t_{travel}^{(1)}$, $t_{travel}^{(2)}$, …, $t_{travel}^{(s)}$) from the exponential distribution ($s$=500). For a given travel time, say $t_{travel}^{(1)}$, we estimate the corresponding time of registration $t_{reg}^{(1)}$ (as $t_{home} + t_{travel}^{(1)}$). Having estimated $t_{reg}^{(1)}$, we construct the corresponding feature vector $X^{(1)}$ (Figure 10D) for use as input for QRF (Figure 10E depicts learning using one feature and one tree). For a given hospital site, feature vectors corresponding to different times of registration are used as inputs for QRF, to generate probabilistic forecasts of waiting times. Travel time estimates are added to the corresponding waiting time estimates, to compute the probability distribution of combined travel and waiting times (Figure 10F). In this example, using second-order stochastic dominance to select between the distributions for the two hospitals, Anna would select the JR hospital. This would lead to an increase in the patient-flow at the JR. Since features are dynamically extracted for each patient, changes in load and congestion from Anna's decision would be reflected in the features for the next incoming patient at the JR. Feedback of patient- and EMT-decisions is thus included during the simulation.

It can be envisaged that providing information on estimates of travel and waiting can influence both patient's and EMT's routing decisions, which can, in turn, affect the load (number of patients) and congestion (waiting times) at an ED. We study this impact by considering the following four alternative decision-making criteria for selecting between the two EDs: (1) shortest distance, (2) lowest travel time

(where travel time was assumed to be the average of the minimum and maximum travel times obtained from Google maps), (3) lowest sum of travel and waiting times (i.e., sum of the travel time used in (2) and a point forecast of the waiting time), and (4) second-order stochastic dominance using the distribution of the sum of travel and waiting times. Out of the above four decision-making criteria, only (4) considers the forecast uncertainty in travel and wait times. Specifically, the rationale of the stochastic dominance criterion is to compare the different CDFs, denoted by, say, $F_{JR}$ and $F_{HG}$, corresponding to the sum of travel and waiting times for the JR and HG, respectively, and select the JR if $\int_{-\infty}^{x}(F_{JR}(t) - F_{HG}(t))dt \geq 0$, and select the HG otherwise. Leshno and Levy (2002) provide a discussion on using stochastic dominance for decision-making. Note that (4) requires the convolution of travel and waiting time distributions, which further highlights the value of estimating the full waiting time distribution in our approach.

**Figure 10.** Schematic diagram illustrates patient decision-making regarding the choice between the JR or HG hospitals. Panel **A** presents the demographics, time of decision, and location postal code. Panel **B** depicts travel routes to the two EDs. Panel **C** shows the exponential distribution of travel times. Panel **D** shows that for each sample of travel time, we have a corresponding estimate for the time of registration and feature vector. Panel **E** depicts a feature vector being used as an input for a regression tree. Panel **F** shows the empirical cumulative distribution function (ECDF) for combined travel and waiting times.




Tables 6 and 7 summarise the simulation results. Table 6 reports the total load resulting at the two hospitals (JR and HG) when each of the four ED selection criteria are used. Table 7 reports the waiting times resulting at the two hospitals. In each table, results for self- and ambulance-arrivals are presented separately. The tables suggest that selecting an ED based solely on the shortest distance criterion will

result in diverting the vast majority of patients (187 self-arrivals out of 214, and all 104 ambulance-arrivals) to the JR (Table 6), which would lead to exceedingly long waiting times at the JR, compared to the HG (Table 7). Considering travel times only would also result in a highly nonuniform spread of load, whereby the HG would experience relatively high attendances and prolonged waiting times. Note that although the vast majority of the simulated postcodes (Supplement, Figure A2) were geographically closer to the JR on average (mean 14.0 miles) compared to the HG (mean 15.9 miles), the overall travel times to the JR were slightly longer (mean 28.9 minutes) than the HG (mean 26.2 minutes). Interestingly, selecting an ED based on just the point estimate of the travel and waiting times, specifically the sum of the travel and waiting times, results in a more uniform spread of patient load across the two hospitals. Encouragingly, making routing decisions based on stochastic dominance, which compares the full distributions of the combined travel and waiting times for the two EDs, results in a more uniform spread of patient load and lowest congestion, whereby no patient ends up waiting for more than 2 hours. Since the JR is a larger hospital with more resources compared to the HG, the model typically ends up diverting more self- and ambulance-arrivals to the JR. The spread of load is similar for the third and fourth criteria in Table 6, whereas selecting an ED based on the probabilistic approach (i.e., stochastic dominance) results in the lowest waiting times as shown in Table 7. As expected, a larger load on the ED seems to be associated with longer waiting times in Table 7. These results suggest that assisting patients in making informed routing decisions can potentially facilitate the uniform spread of load on EDs and help reduce waiting times, which is in broad agreement with previous findings (Dong et al. 2019).

In the absence of actual waiting times for both hospitals for a given patient, point forecasts obtained from the QRF were treated as the true waiting times. Since it was not possible to estimate the time elapsed from arrival at the hospital to registration at the ED (e.g., time spent in the parking area), we treat the time of arrival at the ED as the actual time of registration for simulation. Thus, the numbers quoted in Tables 6 and 7 should be treated with caution and used only for comparative analysis across different criteria. Moreover, in our routing scheme, the ED waiting time estimates are updated when a patient registers at the hospital rather than when a driver selects an ED site from a geographical neighbourhood. Thus, if multiple drivers (either self-arriving patients or EMTs) were to make routing decisions simultaneously, there is a risk that the majority of them may end up selecting the same ED site with lower waiting times (after considering the combined waiting and travel time), which could result in ED congestion. An interesting line of future work on patient routing would be to investigate the situation where drivers are asked to commit to an ED site when the driver starts their journey. This

information could then be used to update waiting time estimates, which would be conveyed to other drivers so they can make a more informed real-time routing decision.

**Table 6.** Number of patients attending the JR and HG hospitals when the choice between the two is based on four alternative selection criteria: shortest distance, lowest mean travel time, lowest sum of mean travel and mean waiting times, and stochastic dominance based on the distribution of the sum of travel and waiting times. Values presented separately for self-arrivals and ambulance-arrivals.

| | **ED Selection Criteria** | | | |
|---|---|---|---|---|
| | Distance | Travel time | Mean Travel + Mean Wait Times | Stochastic Dominance |
| | *Self-arrivals* | | | |
| $n_{JR}$ (low, medium, high) | 187 (9,119,59) | 31 (24,7,0) | 119 (66,52,1) | 118 (88,30,0) |
| $n_{HG}$ (low, medium, high) | 27 (25,2,0) | 183 (9,69,105) | 95 (43,52,0) | 96 (67,29,0) |
| | *Ambulance-arrivals* | | | |
| $n_{JR}$ (low, medium, high) | 104 (4,64,36) | 40 (33,7,0) | 61 (33,28,0) | 57 (41,16,0) |
| $n_{HG}$ (low, medium, high) | 0 (0,0,0) | 64 (1,30,33) | 43 (19,24,0) | 47 (31,16,0) |

*Note*: $n_{JR}$ and $n_{HG}$ denotes total attendance at the JR and HG hospitals, respectively. The number of patients waiting for low (< 45 minutes), medium (46 to 120 minutes), and high (> 120) times are shown in parentheses.

**Table 7.** Waiting times for the JR and HG hospitals when the choice between the two is based on four alternative ED selection criteria: shortest distance, lowest mean travel time, lowest sum of mean travel and mean waiting times, and stochastic dominance based on the distribution of the sum of the travel and waiting times. Values presented separately for self-arrivals and ambulance-arrivals.

| | **ED Selection Criteria** | | | |
|---|---|---|---|---|
| | Distance | Travel time | Mean Travel + Mean Wait Times | Stochastic Dominance |
| | *Self-arrivals* | | | |
| $\mu_{t_{1,JR}}$ ($\sigma_{t_{1,JR}}$) | 108.1 (49.7) | 40.6 (18.4) | 48.4 (21.1) | 39.4 (19.7) |
| $\mu_{t_{1,HG}}$ ($\sigma_{t_{1,HG}}$) | 26.2 (9.4) | 131.2 (53.3) | 50.7 (18.4) | 39.8 (13.4) |
| | *Ambulance-arrivals* | | | |
| $\mu_{t_{1,JR}}$ ($\sigma_{t_{1,JR}}$) | 109.0 (47.7) | 38.1 (17.7) | 52.9 (26.0) | 39.1 (20.3) |
| $\mu_{t_{1,HG}}$ ($\sigma_{t_{1,HG}}$) | - | 132.0 (58.9) | 52.2 (18.8) | 41.9 (15.5) |

*Note*: The mean and standard deviation of waiting time for the JR are denoted by $\mu_{t_{1,\mathrm{JR}}}$ and $\sigma_{t_{1,\mathrm{JR}}}$, respectively. Similar notation is used for the HG hospital.

Currently, patients and EMTs typically use the shortest geographic distance for selecting an ED site. In future, EDs may be able to provide waiting time estimates to either the patients or the EMTs depending on the availability of resources. We thus investigate three routing scenarios to better

understand operational value of using waiting times, whereby: (1) Self-arriving patients use geographic distance and EMTs employ stochastic dominance. (2) Patients use stochastic dominance and EMTs employ geographic distance. (3) Both patients and EMTs use stochastic dominance. To achieve uniform load and lower congestion across EDs, it is imperative to make waiting time estimates accessible for both patients and EMTs (see Supplement Tables 7 and 8).

## 7. Summary and Concluding Remarks

In this study, we propose a machine learning approach using QRF to produce probabilistic forecasts of patient waiting times in an ED. The model utilized a rich set of features that were extracted from detailed patient-level records spanning five years. Rankings of predictor importance suggested that ED workload due to patient volumes and calendar effects were the most salient features. Model evaluation was based on an exhaustive comparison of distributional, quantile, and point forecast accuracy. Encouragingly, QRF convincingly outperformed the empirical benchmarks that are typically used in practice, along with the Q-Lasso and quantile regression methods that have been proposed in the literature for modelling waiting times. The performance of QRF was consistently superior for both hospital sites that we analysed. In addition, we show that the estimates of waiting times can be used to make inferences about the total LOS.

Our findings provide the following operational insights: (1) Compared to selecting an ED based on the shortest geographic distance or lowest travel times, point estimates of ED waiting times allow for more informed routing decisions. (2) To choose an ED, accommodating the uncertainty in waiting times results in a uniform spread of patient load and lower congestion across the two ED sites. This underscores the importance of using uncertainty in forecasts for routing, instead of relying on only a point estimate. (3) To translate the benefits of informed routing decisions for EDs, waiting time estimates should be made available to both patients and EMTs.

Arrivals of low-acuity patients in the ED significantly increase the waiting times for high-acuity patients (Bayati 2017). This is an issue for high-acuity patients that need to be admitted, because delays in admitting a patient to the intensive care unit (ICU) has adverse effects on patient outcomes (Chan et al. 2017). Thus, although we model waiting times of only low-acuity patients in EDs, this work has implications for high-acuity patients and other parts of the hospital, such as ICUs. Given that low-acuity patients are likely to drop out from EDs rather than wait in a crowded room for long hours, especially in times of social distancing, it can be envisaged that publishing waiting times of different service providers could be of particular benefit to patients and EDs.

## Acknowledgements

We are grateful to the Department Editor, Associate Editor, and the three referees for their valuable comments. We also thank Dr Chris Bunch and Aimee Jell for their help in making this work possible.

## References

Anderson, R. T., Camacho, F. T., and Balkrishnan, R. 2007. "Willing to Wait? The Influence of Patient Wait Time on Satisfaction with Primary Care," *BMC Health Services Research* (7).

Ang, E., Kwasnick, S., Bayati, M., Plambeck, E. L., and Aratow, M. 2016. "Accurate Emergency Department Wait Time Prediction," *M&SOM-Manufacturing & Service Operations Management* (18:1), pp. 141-156.

Akşin, Z., Deo, S., Jónasson, J.O., and Ramdas, K. 2021. "Learning from Many: Partner Exposure and Team Familiarity in Fluid Teams," *Management Science* (67:2), pp. 854-874.

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y.N., Tseytlin, Y., Yom-Tov, G. B. 2015. "On Patient Flow in Hospitals: A Data-Based Queueing-Science Perspective," *Stochastic Systems* (5:1), pp. 146-194.

Asaro, P. V., Lewis, L. M., and Boxerman, S. B. 2007. "The Impact of Input and Output Factors on Emergency Department Throughput," *Academic Emergency Medicine* (14:3), pp. 235-242.

Baril, C., Gascon, V., and Vadeboncoeur, D. 2019. "Discrete-Event Simulation and Design of Experiments to Study Ambulatory Patient Waiting Time in an Emergency Department," *Journal of the Operational Research Society* (70:12), pp. 2019-2038.

Batt, R. J., and Terwiesch, C. 2015. "Waiting Patiently: An Empirical Study of Queue Abandonment in an Emergency Department," *Management Science* (61:1), pp. 39-59.

Bayati, M., Kwasnick, S., Luo, D., Plambeck, E. L. 2017. "Low-Acuity Patients Delay High-Acuity Patients in an Emergency Department," *Available at SSRN*).

Ben Taieb S., Taylor, J. W., and Hyndman, R. J. 2021. "Hierarchical Probabilistic Forecasting of Electricity Demand with Smart Meter Data," *Journal of the American Statistical Association* (116:533), pp. 27-43.

Bernstein, S. L., Aronsky, D., Duseja, R., Epstein, S., Handel, D., Hwang, U., McCarthy, M., McConnell, K. J., Pines, J. M., Rathlev, N., Schafermeyer, R., Zwemer, F., Schull, M., Asplin, B. R., Med, S. A. E., and Force, E. D. C. T. 2009. "The Effect of Emergency Department Crowding on Clinically Oriented Outcomes," *Academic Emergency Medicine* (16:1), pp. 1-10.

Boudreaux, E. D., Ary, R. D., Mandry, C. V., and McCabe, B. 2000. "Determinants of Patient Satisfaction in a Large, Municipal Ed: The Role of Demographic Variables, Visit Characteristics, and Patient Perceptions," *American Journal of Emergency Medicine* (18:4), pp. 394-400.

Breiman, L. 2001. "Random Forests," *Machine Learning* (45:1), pp. 5-32.

Buchan, J., Charlesworth, A., Gershlick, B., and Seccombe, I. 2019. "A Critical Moment: NHS Staffing Trends, Retention and Attrition," The Health Foundation.

Carter, E. J., Pouch, S. M., and Larson, E. L. 2014. "The Relationship between Emergency Department Crowding and Patient Outcomes: A Systematic Review," *Journal of Nursing Scholarship* (46:2), pp. 106-115.

Chan, C. W., Farias, V. F., and Escobar, G. J. 2017. "The Impact of Delays on Service Times in the Intensive Care Unit," *Management Science* (63:7), pp. 2049-2072.

Deo, S., and Gurvich, I. 2011. "Centralized Vs. Decentralized Ambulance Diversion: A Network Perspective," *Management Science* (57:7), pp. 1300-1319.

Detmer, D., Bloomrosen, M., Raymond, B., and Tang, P. 2008. "Integrated Personal Health Records: Transformative Tools for Consumer-centric Care," *BMC Medical Informatics and Decision Making* (8:45).

Ding, R., McCarthy, M. L., Desmond, J. S., Lee, J. S., Aronsky, D., and Zeger, S. L. 2010. "Characterizing Waiting Room Time, Treatment Time, and Boarding Time in the Emergency Department Using Quantile Regression," *Academic Emergency Medicine* (17:8), pp. 813-823.

Dong, J., Yom-Tov, E., and Yom-Tov, G. B. 2019. "The Impact of Delay Announcements on Hospital Network Coordination and Waiting Times," *Management Science* (65:5), pp. 1969-1994.

Duan, T., Anand, A., Ding, D.Y., Thai, K.K., Basu, S., Ng, A. and Schuler, A. 2020. "Ngboost: Natural Gradient Boosting for Probabilistic Prediction", In *International Conference on Machine Learning*, pp. 2690-2700.

Epstein, E. S. 1969. "A Scoring System for Probability Forecasts of Ranked Categories," *Journal of Applied Meteorology* (8:6), pp. 985–987.

Friedman, J.H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics* (29:5), pp. 1189–232.

Galarraga, J. E., and Pines, J. M. 2016. "Costs of Ed Episodes of Care in the United States," *American Journal of Emergency Medicine* (34:3), pp. 357-365.

Gaur, V., Kesavan, S., Raman, A., and Fisher, M. L. 2007. "Estimating Demand Uncertainty Using Judgmental Forecasts," *Manufacturing & Service Operations Management* (9:4), pp. 480-491.

Gneiting, T. 2011. "Making and Evaluating Point Forecasts," *Journal of the American Statistical Association* (106:494), pp. 746-762.

Gneiting, T., and Raftery, A. E. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association* (102:477), pp. 359-378.

Goeman, J. J. 2010. "L1 Penalized Estimation in the Cox Proportional Hazards Model," *Biometrical Journal* (52:1), pp. 70-84.

Guo, X., Grushka-Cockayne, Y., and De Reyck, B. 2021. "Forecasting Airport Transfer Passenger Flow using Real-Time Data and Machine Learning," *Manufacturing & Service Operations Management*, forthcoming.

Hastie, T., Tibshirani, R., Friedman, J. H. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: NY: Springer.

He, H. B., and Garcia, E. A. 2009. "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering* (21:9), pp. 1263-1284.

Hing, E., and Bhuiya, F. 2012. "Wait Time for Treatment in Hospital Emergency Departments: 2009," *NCHS Data Brief*:102), pp. 1-8.

Hu, X., Barnes, S., and Golden, B. 2018. "Applying Queueing Theory to the Study of Emergency Department Operations: A Survey and a Discussion of Comparable Simulation Studies," *International Transactions in Operational Research* (25:1), pp. 7-49.

Jouini, O., Aksin, Z., and Dallery, Y. 2011. "Call Centers with Delay Information: Models and Insights," *M&SOM-Manufacturing & Service Operations Management* (13:4), pp. 534-548.

Keskinocak, P., and Savva, N. 2020. "A Review of the Healthcare-Management (Modeling) Literature Published in Manufacturing & Service Operations Management," *M&SOM-Manufacturing & Service Operations Management* (22:1), pp. 59-72.

Leshno, M., and Levy, H. 2002. "Preferred by "All" and Preferred by "Most" Decision Makers: Almost Stochastic Dominance," *Management Science* (48:8), pp. 1074-1085.

Meinshausen, N. 2006. "Quantile Regression Forests," *Journal of Machine Learning Research* (7), pp. 983-999.

Misic, V. V., and Perakis, G. 2020. "Data Analytics in Operations Management: A Review," *M&SOM-Manufacturing & Service Operations Management* (22:1), pp. 158-169.

Morss, R. E., Demuth, J. L., and Lazo, J. K. 2008. "Communicating Uncertainty in Weather Forecasts: A Survey of the US Public," *Weather and Forecasting* (23:5), pp. 974-991.

NHS. 2019. "Handbook to the Nhs Constitution for England."

Rostami-Tabar, B., and Ziel, F. 2020. "Anticipating Special Events in Emergency Department Forecasting," *International Journal of Forecasting* (In Press).

Salari, N., Liu, S., and Shen, Z. J. M. 2022. "Real-time Delivery Time Forecasting and Promising in Online Retailing: When Will Your Package Arrive?," *Manufacturing & Service Operations Management*, forthcoming.

Seiffert, C., Khoshgoftaar, T. M., Hulse, J.V., and Napolitano, A. 2010. "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance," in *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* (40:1), pp. 185-197.

Singer, A. J., Thode, H. C., and Pines, J. M. 2019. "Us Emergency Department Visits and Hospital Discharges among Uninsured Patients before and after Implementation of the Affordable Care Act," *JAMA Network Open* (2:4).

Singh, K. C. D., Scholtes, S., and Terwiesch, C. 2020. "Empirical Research in Healthcare Operations: Past Research, Present Understanding, and Future Opportunities," *M&SOM-Manufacturing & Service Operations Management* (22:1), pp. 73-83.

Singh, K. C. D., and Terwiesch, C. 2012. "An Econometric Analysis of Patient Flows in the Cardiac Intensive Care Unit," *M&SOM-Manufacturing & Service Operations Management* (14:1), pp. 50-65.

Summary, O. 2013. "Waiting Time Policies in the Health Sector Whatworks?." OECD.

Sun, Y., Teow, K. L., Heng, B. H., Ooi, C. K., and Tay, S. Y. 2012. "Real-Time Prediction of Waiting Time in the Emergency Department, Using Quantile Regression," *Annals of Emergency Medicine* (60:3), pp. 299-308.

Tassone, J., Choudhury, S. 2020. "A Comprehensive Survey on the Ambulance Routing and Location Problems," ArXiv, abs/2001.05288.

Taylor, J. W. 2012. "Density Forecasting of Intraday Call Center Arrivals Using Models Based on Exponential Smoothing," *Management Science* (58:3), pp. 534-549.

Ward, M. J., Self, W. H., and Froehle, C. M. 2015. "Effects of Common Data Errors in Electronic Health Records on Emergency Department Operational Performance Metrics: A Monte Carlo Simulation," *Academic Emergency Medicine* (22:9), pp. 1085-1092.

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

Xie, B., and Youash, S. 2011. "The Effects of Publishing Emergency Department Wait Time on Patient Utilization Patterns in a Community with Two Emergency Department Sites: A Retrospective, Quasi-Experiment Design," *International Journal of Emergency Medicine* (4:1), pp. 4-29.

Ye, H., Luedtke, J., and Shen, H. 2019. "Call Center Arrivals: When to Jointly Forecast Multiple Streams?," *Production and Operations Management* (28:1), pp. 27-42.