Forecasting daily patient outflow from a ward having no real-time clinical data

Abstract

Background: Modeling patient flow is crucial in understanding resource demand and prioritization. We study patient outflow from an open ward in an Australian hospital, where currently bed allocation is carried out by a manager relying on past experiences and looking at demand. Automatic methods that provide a reasonable estimate of total next-day discharges can aid in efficient bed management. The challenges in building such methods lie in dealing with large amounts of discharge noise introduced by the nonlinear nature of hospital procedures, and the non-availability of real-time clinical information in wards.

Objective: Our study investigates forecasting total number of next-day discharges from an open ward having no real-time clinical data.

Methods: We compared five popular regression algorithms to model total next-day discharges: (i) auto-regressive integrated moving average (ARIMA) (ii) the auto-regressive moving average with exogenous variables (ARMAX) (iii) k-Nearest Neighbour (kNN) regression (iv) random forest (RF) regression, and (v) support vector regression (SVR). While the ARIMA model relied on past 3 months discharges, nearest neighbor forecasting employed median of similar discharges in the past in estimating next day discharge. Additionally, the ARMAX model used the day of the week and number of patients currently in ward as exogenous variables. For the random forest and support vector regression models, we designed a predictor set of 20 patient features and 88 ward level features.

Results: Our data consisted of 12,141 patient visits over 1826 days. Forecasting quality was measured using Mean Forecast Error, Mean Absolute Error, symmetric Mean Absolute Percentage Error and Root Mean Square Error. When compared to a moving average prediction model, all five models demonstrated superior performance with the random forests achieving 22.7% improvement in Mean Absolute Error, for all days in the year 2014.

Conclusions: In the absence of clinical information, our study recommends using patient level and ward level data in predicting next day discharges. RF and SVR models are able to utilize all available features from such data, resulting in superior performance over traditional auto-regressive methods. An intelligent estimate of available beds in wards plays a crucial role in relieving access block in emergency departments.

Keywords: Patient flow; Discharge planning; Predictive models

Introduction

Demand for healthcare services has become unsustainable [1, 2]. This is largely due to increase in population and life expectancy, escalating costs, increased patient expectations and workforce issues [3]. Despite increased demands, the number of inpatient beds in hospitals has come down by 2% since the last decade [2, 4]. Efficient bed management is key to meeting this rising demand and reducing healthcare costs.

Daily discharge rate can be a potential real-time indicator of operational efficiency [5]. From a ward level perspective, a good estimate of next day discharges will enable hospital staff to foresee potential problems such as changes in number of available beds and changes in number of required staff. Efficient forecasting reduces bed crisis and improves resource allocation. This foresight can help accelerate discharge preparation, which has huge cost on clinical staff and educating patients and family, requiring post-discharge planning [6, 7]. However, studying patient flow from general wards offers several challenges.

Ward level discharges incorporate far greater hospital dynamics that are often non-linear [8]. Accessing real-time clinical information in wards can be difficult because of administrative and procedural barriers, such data may not be available for predictive applications. Since the diagnosis coding is performed post discharge, there is little information about medical condition or variation in care quality in real-time. Also, factors other than patient condition play a role in discharge decisions [5, 9, 10].

The current practice of bed allocation in general wards of most hospitals involves a hospital staff/team, who use past information and experience, to schedule and assign beds [11]. Modern machine learning techniques can be used to aid such decisions and help understand the underlying process. As an example, Figure 1 illustrates a decision tree trained on past discharges and ward occupancy statistics, which models the daily discharge pattern from an open ward in a regional Australian hospital. Although the absence of patient medical information affected forecast performance, the decision rules provide important insight into the discharge process.

Figure 1: Decision tree modelling of total discharges from an open ward from day of the week and ward occupancy (Prev. day Occup) data for five years. The leafs represent total number of patient discharges.

Motivated by this result, we address the open problem of forecasting daily discharges from a ward with no real-time clinical data. Specifically, we compare the forecasting performance of five popular regression models: (i) the classical auto-regressive integrated moving average (ARIMA) (ii) the auto-regressive moving average with exogenous variables (ARMAX) (iii) k-Nearest Neighbour (kNN) regression (iv) random forest (RF) regression, and (v) support vector regression (SVR). Our experiments were conducted on commonly available data from a

recovery ward (Heath Wing 5) in Barwon Health, a regional hospital in Victoria, Australia. The ARIMA and kNN models are built from daily discharges from ward. To account for the seasonal nature of discharges, the ARMAX model included day of the week and ward occupancy statistics. We identified and constructed 20 ward level and 88 patient level predictors to derive the RF and SVR models.

Forecasting accuracy was measured using three metrics on a held out set of 2511 patient visits in the year 2014. When compared to a naive forecasting method of using the mean of last week discharges, we demonstrate through our experiments that (i) using regression methods for forecasting discharge outperforms naive forecasting (ii) SVR and RF models outperforms the auto-regressive methods and kNN (iii) a random forest model derived from 108 features has the minimum error for next-day forecasts.

The significance of our study is in identifying the importance of foreseeing available beds in wards, which could help relieve emergency access block [12].

Related Work
Patient length of stay directly contributes to hospital costs and resource allocation. Long term forecasting in healthcare aims to model bed and staffing needs over a period of months to years. Cote and Tucker categorize the common methods in healthcare demand forecasting as: percent adjustment, 12-month moving average, trendline and seasonalized forecast [13]. While each of these methods are built from historical demand, seasonalized forecasting provides more realistic results are it takes into account the seasonal variations and trends in the data. Mackay and Lee [3] advise modeling the patient flow in healthcare institutions for tactical and strategic forecasting. To this end, compartmental modeling [14, 15], queuing models [16, 17] and simulation models [17-20] have been applied to analyze patient flow. To understand long term patient flow, studies analyze metrics such as bed occupancy [3, 8, 14, 19, 21, 22], patient arrivals [23] and individual patient length of stay [19, 24-27].

On the other hand, our work implements short-term forecasting. The short-term forecasting methods are concerned with hourly and daily forecasts from a single unit in a care environment. The most popular unit of interest is the emergency or acute care department, since this is often a key performance indicator (KPI) metric in assessing quality of care [28, 29].

Time-series and smoothing methods
When looking at discharges as time series, auto regressive moving average models are the most popular [30-32]. Exponential smoothing techniques have also been used to forecast monthly [33] and daily patient flow [34].
Jones et al. used the classical autoregressive integrated moving average (ARIMA) to forecast daily bed occupancy in emergency department of a European hospital [30]. The model with included seasonality terms demonstrated reasonable performance to predict bed occupancy. The authors speculated whether nonlinear forecasting

techniques could improve over ARIMA. A recent study confirmed the effectiveness of this forecasting technique in a US hospital setting [35]. ARIMA models were also successfully used to forecast the number of occupied beds during a SARS outbreak in a Singapore hospital [36]. A recent study used patient attendances in a paediatric emergency department to model daily demand using ARIMA [37].

Jones et al. [34] compared the ARIMA mode with exponential smoothing and artificial neural networks to forecast daily patient volumes in emergency department. The study revealed no single model to be superior, and concluded that seasonal patterns play a major role in daily demand.

Simulation Methods

Modelling using simulation is typically used to study the behavior of complex systems. An early work in investigated the effects of emergency admissions on daily bed requirements in acute care, using discrete-event stochastic simulation modelling [38]. Sinreich and Marmor [39] proposed a guide for building a simulation tool based on data from emergency departments of 5 Israeli hospitals. Their method analyzed the flow of patients clustered into eight types along with time elements. The simulation demonstrated that patient processes are better characterized by type of the patients, rather than specific hospitals visited. Yeh and Lin used a simulation model to characterize patient flow through a hospital emergency department, and reduced waiting times using a genetic algorithm [40]. A similar experiment was carried out in a geriatric department using a combination of discrete event simulation and queueing model to analyze bed occupancy [19].

Regression for forecasting

Regression models analyse the relationship between the forecasted variable and features in the data. Linear regression that encoded monthly variations were used to forecast patient admissions over a six months horizon, and outperformed quadratic and auto-regressive models [41]. Another study used clustering and PCA to find significant predictors from patient data to model emergency length-of-stay using linear regression [42]. A nonlinear approach using regression trees was proposed in forecasting patient admissions which demonstrated superior performance over a neural net framework [43].

Barnes et al. used 10 predictors to model real-time inpatient length of stay in a 36 bed unit using a random forest model [24].

Nonlinear regression is better suited to model the changing dynamics of patient flow. To characterize the outflow of patients from the ward, we resort to regression using random forest, k-Nearest Neighbour and support vector regression. In the area of pattern recognition, k-nearest neighbours (kNN)[44] are the most effective method that exploits repeated patterns. The kNN algorithm has been successfully applied to forecast to histogram time series in financial data [45]. The non-parametric regression using kNN has been successfully demonstrated for short term traffic forecasting [46, 47], electricity load forecasting [48, 49] . However, kNN regression has not been studied for patient flow.

Support Vector regression (SVR) is also a powerful and popular regression technique that uses kernel functions to map features into a higher dimensional space to perform linear regression. Though this technique has not seen much application in medical forecasting, support vector machines have been successful in financial market prediction, electricity forecasting, business forecasting and reliability forecasting [50].

Apart from the standard auto-regressive methods, we use kNN, random forests and SVR in forecasting next-day discharges. Since discharge patterns repeat over time, k-Nearest Neighbour regression can be applied to search for a matching pattern from past discharges. Random forests and SVR regression are powerful modelling techniques requiring minimum tuning to effectively handle nonlinearity in the hospital processes.

Recently, random forest forecasting was used to predict total patient discharges from a 36 bed unit in an urban hospital [24]. Apart from four demographic and two timing predictors, this study used three clinical predictors for patients: (i) reason for visit: identified by a physician and recorded using ICD-9 diagnosis codes [51] (ii) observation status: assigned to patients for monitoring purpose, and (iii) pending discharge location. Total number of discharges were estimated from aggregate of individual patient length-of-stay.

The absence of real-time clinical information in our data makes calculating patient length-of-stay impossible. Instead, we resort to modelling next day discharges by observing previous discharge patterns, and examining demographics and flow characteristics in the ward.

Methods

Data

Our study used retrospective data collected from a recovery ward in Barwon Health, a large public health provider in Victoria, Australia serving about 350,000 residents. Ethics approval was obtained from the Hospital and Research Ethics Committee at Barwon Health (number 12/83) and Deakin University. The total number of available beds depended on the number of staff assigned to the ward. On average, the ward had 36 staffed beds, but fluctuated between 20 and 80 beds with varying patient flow. The physicians in the ward had no teaching responsibilities.

Table 1: Tables in hospital database used in our data collection

Patients	
1	Patient ID
2	Age
3	Gender
Ward Stay	
1	Admission ID
2	Name of the ward
3	Time (entry, exit)
4	Bed ID
Admissions	
1	Patient ID
2	Admission ID
3	Time (admit, discharge)
4	Patient Class (21 categories)
5	Admission type (7 categories)
6	Reference to speciality (49 categories)

The data for our study came from three tables in the hospital database, as shown in Table 1. Additional real-time data that described patient condition or disease progression were unavailable, since diagnosis coding using medical codes is done after discharge. Patient flow was collected for a period of 4 years. Using the admission and discharge times for each patient, we calculated the daily discharges from our ward in study. A total of 12,141 patients were admitted into the ward with a median discharge of 8 patients per day from January 1, 2010 to December 31, 2014. Table 2 summarizes the main characteristics of our data.

Table 2. Cohort Details

Total patient visits	12,141
Unique patients	10,610
Length of stay: mean, median, IQR	4.26, 3, 5
Discharges per day: mean, median, IQR	8.7, 8, 5
Admissions per day: mean, median, IQR	8.6, 8, 5
Mean ward occupancy, IQR	30.9, 4
Gender	54.8% Female
Age: mean, median	66, 63.23

A time series decomposition of our data revealed strong seasonal variations and high non-linearity in daily discharge patterns. There was a defined weekly pattern - discharge from ward peaked on Fridays and dropped significantly on weekends (see Figure 2). This seasonal nature is in tune with previous studies [9, 32]. Aggregating the daily discharges into a monthly time series revealed defined monthly patterns

(see Figure 3). The data displayed no significant trend. In addition, the daily discharge pattern was found to be highly nonlinear. Our forecasting methods must be able to handle such data dynamics.

Figure 2: Mean admissions and discharges per day from ward

Figure 3 Time series of monthly discharges from ward

Forecasting Methods

We describe the following diverse methods that are applicable to forecasting under complex data dynamics: (i) auto-regressive integrated moving average (ARIMA), (ii) auto-regressive moving average with exogenous variable (ARMAX) (iii) forecasting using k-Nearest Neighbour (kNN) discharge patterns and (iv) random forest (RF) (v) support vector regression (SVR). Auto regressive methods model the temporal linear correlation between nearby data points in the time-series. Nearest patterns lifts this linearity assumption and assumes that short-periods form repeated patterns. Finally, random forest and support vector regression look for a nonlinear functional relationship between the future outcomes and descriptors in the past.

Auto-regressive integrated moving average (ARIMA)

Time-series forecasting methods can analyse the pattern of past discharges and formulate a forecasting model from underlying temporal relationships [52]. Such models can then be used to extrapolate the discharge time series into the future. Auto regressive integrated moving average (ARIMA) models are widely used in time-series forecasting. Their popularity can be attributed to ease of model formulation and interpretability [53]. ARIMA models look for linear relationships in the discharge sequence to detect local trends and seasonality. However, such relationships can change over time. ARIMA models are able to capture these changes and update themselves accordingly. This is done by combining Autoregressive (AR) and moving average (MA) models. Autoregressive models formulate discharge at time $t = y_t$, as a linear combination of previous p discharges. On the other hand, moving averages models characterize y_t as linear combination of previous q forecast errors. For ARIMA model, the discharge time series is made stationary using differencing. Let $\emptyset$ be autoregressive parameters, θ be moving average parameters and ϵ be the forecast errors. We can define and ARIMA model as:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t - \sum_{i=1}^{q} \theta_q \, \epsilon_{t-i}$$

where μ is a constant. By varying p and q, we can generate different models to fit the data. Box Jenkins method [54] provides a well-defined approach for model identification and parameter estimation. In our work, we choose the auto.arima() function from the forecast package [55] in R [56] to automatically select the best model.

Auto-regressive moving average with Exogenous Variables (ARMAX)

Dynamic regression techniques allows adding additional explanatory variables, like day of the week and number of current patients in the ward, to autoregressive models. The Auto-regressive Moving Average with Exogenous Variables (ARMAX) modifies ARIMA model by including depending external variable x_t at time t as follows:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t - \sum_{i=1}^{q} \theta_q \epsilon_{t-i} + \beta x_t$$

We model x_t using features from the hospital database.

Detecting Discharge Patterns using k-Nearest Neighbours

The kNN algorithm takes advantage of the locality in data space. We assume the next day discharge depends on the discharges happening in previous d days. Using kNN principles, we can do a regression to forecast the next day discharge. Let y_d represent number of discharges on the current day: d. To forecast the next day discharge: y_{d+1}, we look at the discharges over the past p days as: disch_vec $= [y_{d-p} : y_d]$. Using Euclidean distance metric, we find k closest matches to disch_vec from the training data. An estimate of next day discharge: $\hat{y}_{d+1}$, is calculated as a measure of the next day discharges of the k matched patterns: $(y_{\text{match}})_i$ $i \in (1 : k)$. As example, Figure 4 shows an example of k-Nearest Neighbour based forecasting. Here, disch_vec in red $[y_{d-7} : y_d]$ results in three matches from the training data. For simplicity, we have plotted the matched patterns alongside disch_vec, though they had occurred in the past. The next day forecast $\hat{y}_{d+1}$ becomes a measure of $(y_{\text{match}})_i$, where $(y_{\text{match}})_i$ $i \in (1 : 3)$ is the $(d + 1)^{th}$ term of each of the matched patterns [57].

Figure 4 k-Nearest Neighbour forecasting example with k=3 and p=7

One popular method of calculating $\hat{y}_{d+1}$ is by minimizing the weighted quadratic loss:

$$\hat{y}_{d+1} = \min_y \sum_{i=1}^{k} w_i ((y_{\text{match}})_i - y)^2$$

$$= \sum_{i=1}^{k} w_i (y_{\text{match}})_i$$

where w_i takes values between 0 and 1, with $\sum_{i=1}^{k} w_i = 1$. However, there are two main drawbacks making it less desirable for our data. First, the quadratic loss is sensitive to outliers. Second, a robust estimate of $\{w_i\}$ becomes difficult.
Our data contains significant noise, causing large variations in next day forecasts of the k matched patterns. We illustrate this problem in Figure 5. For a given day, kNN regression returns 125 matched patterns. The next day forecasts from each k=125

patterns displayed significant variations. In such scenario, we resort to estimating $\hat{y}_{t+1}$ by minimizing the following robust loss:

$$\hat{y}_{t+1} = \min_y \sum_{i=1}^{k} |(y_{\text{match}})_i - y|$$

$$= \text{median}(y_{\text{match}})_{i=1 \text{ to } k}$$

Figure 5: Scatterplot of next day forecast using kNN for a given day. X-axis represents each matched nearest neighbor pattern. Y-axis represents the next day forecast of that matched pattern.

Random Forest

In this approach, we assume the next-day discharge as a function of historical descriptor vector: x. We use each day in the past as a data point, where the next-day discharge is the outcome, and the short period prior to the discharge are used to derive descriptors x . The random forest used in this paper is currently one of the most powerful methods to model the function $y = f(x)$ [58, 59]. A random forest is an ensemble of regression trees. A regression tree approximates a function $f(x)$ by recursively partitioning the descriptor space. At each region R_p, the function is approximated as:

$$f(x) = \frac{1}{|R_p|} \sum_{x_j \in R_p} y_j$$

where $|R_p|$ is the number of data point falling in region R_p. The random forest creates a diverse collection of random trees by varying the subsets of data points to train the trees and the subsets of descriptors at each step of space partitioning. The final outcome of random forest is an average of all trees in the ensemble. Since tree growing is a highly adaptive process, it can discover any nonlinear function to any degree of approximation if given enough training data. However, the flexibility makes regression tree prone to overfitting, that is, the inability to generalize to unseen data. This requires controlling the growth by setting the number of descriptors per partitioning step, and the minimum size of region R_p.

The voting leads to great benefits: reduce the variations per tree. The randomness helps combat against overfitting. There is no assumption about the distribution of data, or the form of the function (x) . There is controllable quality of fits.

Support Vector Regression

The historical descriptor vector x, used in the random forest model can also be used to build a Support Vector Regression (SVR) model [60]. Given the set of data $\{(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)\}$, where each $x_i \in R^m$ denotes the input descriptor for the corresponding next day forecast $y_i \in R^1$, a regression function takes the form: $\hat{y}_i = f(x_i)$. Support vector regression works by (i) mapping the input space of

x_i into a higher dimensional space using a non-linear mapping function: $\boldsymbol{\phi}$ (ii) performing a linear regression in this higher dimensional space. In general, we can express the regression function as:

$$f(\boldsymbol{x}) = \left(w\boldsymbol{\phi}(\boldsymbol{x})\right) + b$$

where, $w \in R^m$ is the weights and $b \in R^1$ is the bias term. Vapnik [60] proposed the ϵ-insensitive loss function for SVR, which takes the form:

$$L_\epsilon(f(\boldsymbol{x}) - \boldsymbol{y}) = \begin{cases} |f(\boldsymbol{x}) - \boldsymbol{y}|, & |f(\boldsymbol{x}) - \boldsymbol{y}| \geq \epsilon \\ 0, & \text{otherwise} \end{cases}$$

The loss function L_ϵ tolerates errors that are smaller than the threshold: ϵ, resulting in a "tube" around the true discharge values.

Model parameters can be estimated by minimizing the following cost function:

$$R = C \times \frac{1}{n} L_\epsilon(f(\boldsymbol{x}) - \boldsymbol{y}) + \frac{1}{2} \|w\|^2$$

where C is a constant that penalizes error in training data. In our work, we use an RBF kernel [61] for mapping our input data to higher dimensional feature space. RBF kernels are a good choice for fitting our nonlinear discharge pattern because of its ability to map the training data to an infinite dimensional space, and easy implementation. The solution to the dual formulation of SVR cost function is detailed in [60, 62].

Experiments

We extracted all data from the database tables (as in Table 1.) for our ward in study. Patient flow was analyzed for a period of five years. We formatted our data as a matrix where each row corresponds to a day and each column represents a feature (descriptor). Two main groups of features were identified: (i) ward level (ii) patient level. Our feature creation process resulted in 20 ward level and 88 patient level predictors, as listed in Table 3. The ward level descriptor: trend of next day discharge, was calculated by fitting a locally weighted polynomial regression [63] from past discharges. An example of this regression fitting is shown in Figure 6.

Figure 6: An example of the discharge trend, as derived from a locally weighted polynomial regression model

Table 3: Features constructed from ward data in hospital database. The random forest and SVR models used the full set of features. ARMAX model used seasonality and occupancy. All other models were derived from daily discharges.

Ward level Predictors	
Seasonality	current day-of-week, current month
Trend	calculated using locally weighted polynomial regression from past discharges on the same weekday
Admissions	Number of admissions during past 7 days
Discharges	Number of discharges during past 7 days, number of discharges in previous 14th day and 21st day
Occupancy	ward occupancy in previous day
Patient level Predictors	
Admission type	5 categories
Patient Referral	49 categories
Patient Class	21 categories
Age Category	8 categories
Number of wards visited	4 categories
Elapsed length of stay	Calculated daily for each patient in the ward

Evaluation Protocol

Our training and testing sets are separated by time. This strategy reflects the common practice of training the model using data in the past and applying it on future data. Training data consisted of 1460 days from January 1, 2010 to December 31, 2013. Testing data consisted of 365 days in the year 2014. The characteristics of the training and validation cohort are shown in Table 4. The majority of stays were short, around 65% of patients stayed for less than 5 days.

Table 4 Characteristics of training and validation cohorts.

	Training (2010-2013)	Testing (2014)
Total days	1460	365
Mean discharges per day	8.47	9.17
Number of admissions	9630	2511
Gender		
Male	4329 (44.9%)	1135 (45.2%)
Female	5301 (55.1%)	1376 (54.8%)
Mean age (years)	63.65	61.62
Length of Stay:		

| 1-4 days | 6377 (66.22%) | 1636 (65.15%) |
| 5 or more days | 3253 (33.78%) | 875 (34.85%) |

Baseline Forecasting

The current hospital strategy involves using past experience to foresee available beds. To compare the efficiency of our proposed approaches, we model the following baselines:

- Naive Forecasting using the last day of week discharge: Since our data was found to have defined weekly patterns, we model the next day discharge as the number of discharges for the same day during previous week.
- Naive Forecasting using mean of last week discharges: To better model the variation and noise in weekly discharges, we model the next day discharge as the mean of discharges during previous 7 days.
- Naive Forecasting using mean of last 3 week discharges: To account for the monthly and weekly variations in our data, we use mean of daily discharges over the past 3 weeks to model the next day discharge.

Measuring Forecast Performance

We compare the next day forecasts of our proposed approaches with the baseline methods on the measures of mean forecast error, mean absolute error, symmetric mean absolute percentage error and root mean square error [64, 65]. If y_t is the measured discharge at time t, f_t is the forecasted dishcharge at time t, we can define the following:

- Mean Forecast Error (MFE): is used to gauge model bias and is calculated as:

$$MFE = \text{mean}(y_t - f_t)$$

  For an ideal model, MFE = 0. If MFE > 0, the model tends to under-forecast. When MFE < 0, the model tends to over-forecast.

- Mean Absolute Error (MAE): is the average of unsigned errors.

$$MAE = \text{mean}|y_t - f_t|$$

  MAE indicates the absolute size of the errors.

- Root Mean Square Error (RMSE) is a measure of the deviation of forecast errors. It is calculated as:

$$RMSE = \sqrt{\text{mean}(y_t - f_t)^2}$$

  Due to squaring and averaging, large errors tend to have more influence over RMSE. In contrast, individual errors are weighted equally in MAE. There has

been much debate on the choice of MAE or RMSE as an indicator of model performance [66, 67]

- Symmetric Mean Absolute Percentage Error (sMAPE): is scale independent and hence can be used to compare forecast performance between different data series. It overcomes two disadvantages of Mean Absolute Percentage Error (MAPE) namely, (i) the inability to calculate error when the true discharge is zero (ii) heavier penalties for positive errors than negative errors. sMAPE is a more robust estimate of forecast error and is calculated as:

$$\text{sMAPE} = \text{mean}\left(200\frac{|y_t - f_t|}{y_t + f_t}\right)$$

However, sMAPE ranges from -200% to 200%, giving it an ambiguous interpretation [68].

Results

Model Performance

In this section, we describe the results of comparing our different forecasting methods. The model parameters for kNN forecast, random forest and SVR model were tuned to minimize forecast errors.

For kNN regression, the optimum value of pattern length: d and number of nearest neighbours: k, was obtained by analysing forecast RMSE for values $d \in (1,100)$ and $k \in (5,1000)$. Minimum RMSE of 3.77 was obtained at $d = 70$ and $k = 125$.

The SVR parameters C (penalty cost) and ϵ (amount of allowed error) were determined by choosing the best value from a grid search, that minimized the model RMSE. Similarly, the optimum number of variables in building each node of the random forest was chosen by examining its effect on minimising the out-of-bag estimate.

We compared the naive forecasting methods with our proposed five models using MFE, MAE, RMSE and sMAPE. The results are summarized in Table 5, while Figure 7 compares the distribution of actual discharges with different model forecasts.

Table 5. Forecast accuracy of different models

Model	MFE	MAE	sMAPE	RMSE	
Naive Forecast					
using discharge from last weekday	0.03	3.81	45.70 %	4.95	
using mean of last week discharges	0.02	3.57	41.68 %	4.42	
using mean of last 3 weeks discharges	0.04	3.44	40.14%	4.34	

					MAE Improv over Naive
ARIMA	0.06	3.27	38.32 %	4.15	4.9 %
ARMAX	-0.01	2.99	34.86 %	3.84	13.1 %
kNN	1.09	2.88	34.92 %	3.77	16.3 %
SVR	0.73	2.75	32.88%	3.64	20.1 %
Random Forest	0.44	2.66	31.86 %	3.49	22.7 %

Figure 7 Comparison of actual and forecasted discharges from ward for each day in 2014

The naive forecasts are unable to capture all variations in the data, and resulted in the maximum error when compared to other models.

The variations in seasonality and trend are better captured in ARIMA and ARMAX models. The time-series consisting of past 3 months discharges were used to generate the next day discharge forecast. The ARMAX model also included the day of week and ward occupancy as exogenous variables, which resulted in better forecast performance over ARIMA.

Interestingly, kNN was more successful than ARIMA and ARMAX in capturing the variations in discharge, demonstrating about 3 % improvement in MAE, when compared to ARMAX. However, the kNN model tends to under-forecast (MFE = 1.09), possibly because of resorting to median values for forecast. In comparison, random forest and SVR forecast models demonstrated better performance. This can be expected, since they are derived from all the 108 features. However, random forest demonstrated a relative improvement of 3.3 % in MAE over SVR model (see Table 5). When looking at forecast errors for each day of week, random forest model confirmed better performance, as shown in Figure 8.

The process of support vector regression with RBF kernel maps all data into a higher dimensional space. Hence, the original features responsible for forecast cannot be recovered, and the model acts as a black box. Alternatively, random forest algorithm returns an estimate of importance for each variable for regression. Examining the features with high importance could give us a better understanding of the discharge process.

Figure 8 Forecast error in predicting each day of week in 2014

Feature Importance in Random Forest model
The features in random forecast model were ranked on importance scores. The top 10 significant features are described as follows. The day of week for the forecast proved to be the most important feature. Other features were: number of patients in the ward during the day of forecast, the trend of discharges measured using locally weighted polynomial regression, number of discharges in past 14th day, number of discharges in past 21st day, number of patients who had visited only one

previous ward, the number of males in the ward, number of patients labelled as: "Public standard", and current month of forecast.

Discussion and Conclusion

Improved patient flow and efficient bed management is key to counter escalating service and economic pressures in hospitals. Predicting next day discharges is crucial, but has been seldom studied for general wards. When compared to emergency and acute care wards, predicting next day discharges from a general ward is more challenging because of the non-availability of real-time clinical information. The daily discharge pattern is seasonal and irregular. This could be attributed to management of hospital processes such as ward rounds, inpatient tests, and medication. The nonlinear nature of these processes contribute to unpredictable length of stay even in patients with similar diagnosis.

Typically, for open wards, a floor manager uses previous experience to foresee the number of available beds. In this paper, we attempt to model total number of next day discharges using five methods. We have compared the forecasting performance using MAE, RMSE and sMAPE. Our predictors are extracted from commonly available data in the hospital database. While the k-nearest neighbour method is simple to implement, requiring no special expertise, software packages for other models are available for all common platforms. These models can be implemented by the analytics staff in hospital IT department and can be easily integrated into existing health information systems.

Findings

In our experiments, forecast based on random forest model outperformed all other models. Forecasting error rate is 31.9% (as measured by sMAPE) which is in the same ballpark as the recent work of [24], though we had no real-time clinical information. A random forest model makes minimum assumptions about the underlying data. Hence it is the most flexible, and at the same time, comes with great overfitting control. Similarly, support vector regression also demonstrated superior performance, compared to the auto-regressive and kNN models. The RBF kernel maps the features into a higher dimensional space during the regression process. Hence the physical meaning of the features are lost, making it difficult to interpret the model. Finally, random forests and support vector regression are able to handle more features. This extra information in the form of patient demographics and past admission and discharge statistics contributed to improve the predictive performance when compared to other models.

The kNN regression also performed well as it assumes only the locality in the data. But it is not adaptive, and thus less flexible in capturing complex patterns. The kNN regression assumes similar patterns in past discharges extrapolate to similar future discharge, which is not true for daily discharges from ward. ARMAX model outperformed the traditional ARIMA forecasts since it incorporated seasonal information as external regressors. As expected, a naive forecast of using the median of past discharges performed worst.

We noticed a weekly pattern (Figure 2) and monthly pattern (Figure 3) in discharges from the ward. Other studies have also confirmed that discharges peak on Friday and drop during weekends [5, 9, 10]. This "weekend effect" could be attributed to shortages in staffing, or reduced availability of services like sophisticated tests and procedures [10, 69]. This suggests discharges are heavily influenced by administrative reasons and staffing.

Feature importance score from a random forest model helps in identifying the features contributing to the regression process. The day of forecast proved to be one of the most important features in the random forest model. Other important features included trend based on non-linear regression of past weekdays, number of discharges in the past days, ward occupancy in previous day, number of males in the ward, and number of general patients in ward.

When looking at for each day of the week, the RF and SVR model consistently outperformed other models. Sundays and Thursdays proved to be the easiest to predict for all models (Figure 8). This can be expected since these days had the least variation in our data. Fridays proved to be the most difficult to forecast. Retraining the random forest model by omitting "day of the week" increased the forecast error by 1.39% (as measured by sMAPE).

Nonlinear dynamics of patient flow

Patient length of stay is inherently variable, partly due to the complex non-linear structure of medical care [8]. The number of discharges from a ward is strongly related to the length of stay of the current patients in the ward. Hence, the variability in ward level discharges is compounded by the variability in individual patient length of stay. In our study, the daily discharge pattern from ward shows great variation for each day of week. Apart from patient level details, we believe that a knowledge of hospital policies is also required to capture such non-linearity.

Practical Significance

In our study, we were able to validate that the weekend patterns affect discharges from a general ward. The random forest model was able to give a reasonable estimate of number of next day discharges from the ward. Clinical staff can use this information as an aid to decisions regarding staffing and resource utilization. This foresight can also aid discharge planning such as communication and patient transfer between wards, or between hospitals.

An estimate of number of free beds can also help reduce ED boarding time and improve patient flow [12, 23]. ED boarding time is the time spent by a patient in emergency care when a bed is not available in the ward. ED boarding time severely reduces the hospital efficiency. High bed occupancy in ward directly contributes to ED overcrowding [70]. In our data, 42.81% of patients were admitted from the emergency care. An estimate of daily forecasts can be helpful in deciding the number of beds in wards to ease patient flow.

Study limitations

We acknowledge the following limitations in our study. First, we focused only on a single ward. However, it was a ward with different patient types, and hence the results could be an indication for all general wards. Second, we did not use patient clinical data to model discharges. This was because clinical diagnosis data was available only for 42.81% of patients who came from emergency. In a general ward, clinical coding is not done in real-time. However, we believe that incorporating clinical information to model patient length of stay could improve forecasting performance. Third, we did not compare our forecasts with clinicians/managing nurses. Finally, our study is retrospective. However, we have selected prediction period separated from development period. This has eliminated possible leakage and optimism.

Conclusion

This study set out to model patient outflow from an open ward with no real-time clinical information. We have demonstrated that using patient level and ward level features in modelling forecasts outperforms the traditional auto-regressive methods. Our proposed models are built from commonly available data and hence could be easily extended to other wards. By supplementing patient level clinical information when available, we believe that the forecasting accuracy of our models can be further improved.

Acknowledgements

Conflicts of Interest

None declared.

Abbreviations

ARIMA: Auto-Regressive Intensive Moving Average
ARMAX: Auto-Regressive Moving Average with exogenous variables
ED: Emergency Department
kNN: k-Nearest Neighbour
MAE: Mean Absolute Error
MAPE: Mean Absolute Percentage Error
MFE: Mean Forecast Error
RF: Random Forest
RMSE: Root Mean Square Error
sMAPE: Symmetric Mean Absolute Percentage Error
SVR: Support Vector Regression

References

1. Kalache, A. and A. Gatti, *Active ageing: a policy framework.* Advances in gerontology, 2002. **11**: p. 7-18.
2. OECD, *A Disease-based Comparison of Health Systems.* 2003.
3. Mackay, M. and M. Lee, *Choice of models for the analysis and forecasting of hospital beds.* Health Care Management Science, 2005. **8**(3): p. 221-230.
4. Alijani, A., et al., *Instrument for objective assessment of appropriateness of surgical bed occupancy: validation study.* BMJ, 2003. **326**(7401): p. 1243-1244.
5. Wong, H.J., et al., *Real-time operational feedback: daily discharge rate as a novel hospital efficiency metric.* Quality and Safety in Health Care, 2010. **19**(6): p. 1-5.
6. Connolly, M., et al., *Discharge preparation: Do healthcare professionals differ in their opinions?* Journal of interprofessional care, 2010. **24**(6): p. 633-643.
7. Connolly, M., et al., *Systems and people under pressure: the discharge process in an acute hospital.* Journal of clinical nursing, 2009. **18**(4): p. 549-558.
8. Harper, P.R. and A. Shahani, *Modelling for the planning and management of bed capacities in hospitals.* Journal of the Operational Research Society, 2002: p. 11-18.
9. Wong, H., et al., *How much do operational processes affect hospital inpatient discharge rates?* Journal of public health, 2009: p. fdp044.
10. van Walraven, C. and C.M. Bell, *Risk of death or readmission among people discharged from hospital on Fridays.* Canadian Medical Association Journal, 2002. **166**(13): p. 1672-1673.
11. Daniels, M.J., M.E. Kuhl, and E. Hager. *Forecasting Hospital Bed Availability Using Simulation and Neural Networks.* in *IIE Annual Conference. Proceedings.* 2005. Institute of Industrial Engineers-Publisher.
12. Luo, W., et al., *Estimating the intensity of ward admission and its effect on emergency department access block.* Statistics in medicine, 2013. **32**(15): p. 2681-2694.
13. Cote, M.J. and S.L. Tucker, *Four Methodologies to Improve Healthcare Demand Forecasting*, in *Healthcare Financial Management.* 2001. p. 54.
14. McClean, S. and P.H. Millard, *A decision support system for bed-occupancy management and planning hospitals.* Mathematical Medicine and Biology, 1995. **12**(3-4): p. 249-257.
15. McClean, S.I. and P.H. Millard, *A three compartment model of the patient flows in a geriatric department: a decision support approach.* Health care management science, 1998. **1**(2): p. 159-163.
16. Gorunescu, F., et al., *A queueing model for bed-occupancy management and planning of hospitals.* Journal of the Operational Research Society, 2002. **53**(1): p. 19-24.
17. Mills, T., *A mathematician goes to hospital.* Australian Mathematical Society Gazette, 2004. **31**(5): p. 320-327.
18. Costa, A., et al., *Mathematical modelling and simulation for planning critical care capacity*.* Anaesthesia, 2003. **58**(4): p. 320-327.

19. El-Darzi, E., et al., *A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department.* Health Care Management Science, 1998. **1**(2): p. 143-149.

20. Hoot, N.R., et al., *Forecasting emergency department crowding: a discrete event simulation.* Annals of emergency medicine, 2008. **52**(2): p. 116-125.

21. Mackay, M., *Practical experience with bed occupancy management and planning systems: an Australian view.* Health Care Management Science, 2001. **4**(1): p. 47-56.

22. Gorunescu, F., S.I. McClean, and P.H. Millard, *Using a queueing model to help plan bed allocation in a department of geriatric medicine.* Health Care Management Science, 2002. **5**(4): p. 307-312.

23. Peck, J.S., et al., *Predicting Emergency Department Inpatient Admissions to Improve Same-day Patient Flow.* Academic Emergency Medicine, 2012. **19**(9): p. E1045-E1054.

24. Barnes, S., et al., *Real-time prediction of inpatient length of stay for discharge prioritization.* Journal of the American Medical Informatics Association, 2015: p. ocv106.

25. Levin, S.R., et al., *Real-time forecasting of pediatric intensive care unit length of stay using computerized provider orders.* Critical care medicine, 2012. **40**(11): p. 3058-3064.

26. Clark, D.E. and L.M. Ryan, *Concurrent prediction of hospital mortality and length of stay from risk factors on admission.* Health services research, 2002. **37**(3): p. 631-645.

27. Marshall, A., C. Vasilakis, and E. El-Darzi, *Length of stay-based patient flow models: recent developments and future directions.* Health Care Management Science, 2005. **8**(3): p. 213-220.

28. Kulinskaya, E., D. Kornbrot, and H. Gao, *Length of stay as a performance indicator: robust statistical methodology.* IMA Journal of Management Mathematics, 2005. **16**(4): p. 369-381.

29. Lindsay, P., et al., *The development of indicators to measure the quality of clinical care in emergency departments following a modified-delphi approach.* Academic emergency medicine, 2002. **9**(11): p. 1131-1139.

30. Jones, S.A., M.P. Joy, and J. Pearson, *Forecasting demand of emergency care.* Health care management science, 2002. **5**(4): p. 297-305.

31. Littig, S.J. and M.W. Isken, *Short term hospital occupancy prediction.* Health care management science, 2007. **10**(1): p. 47-66.

32. Lin, R.-C., K.S. Pasupathy, and M.Y. Sir, *Estimating Admissions and Discharges for Planning Purposes-Case of an Academic Health System.* Advances in Business and Management Forecasting, 2011. **8**: p. 115.

33. Lin, W.T., *Modeling and forecasting hospital patient movements: Univariate and multiple time series approaches.* International Journal of Forecasting, 1989. **5**(2): p. 195-208.

34. Jones, S.S., et al., *Forecasting daily patient volumes in the emergency department.* Academic Emergency Medicine, 2008. **15**(2): p. 159-170.

35. Schweigler, L.M., et al., *Forecasting models of emergency department crowding.* Academic Emergency Medicine, 2009. **16**(4): p. 301-308.

36.   Earnest, A., et al., *Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore.* BMC Health Services Research, 2005. **5**(1): p. 36.

37.   Kadri, F., et al., *Time series modelling and forecasting of emergency department overcrowding.* Journal of medical systems, 2014. **38**(9): p. 1-20.

38.   Bagust, A., M. Place, and J.W. Posnett, *Dynamics of bed use in accommodating emergency admissions: stochastic simulation model.* Bmj, 1999. **319**(7203): p. 155-158.

39.   Sinreich, D. and Y. Marmor, *Emergency department operations: the basis for developing a simulation tool.* IIE transactions, 2005. **37**(3): p. 233-245.

40.   Yeh, J.-Y. and W.-S. Lin, *Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department.* Expert Systems with Applications, 2007. **32**(4): p. 1073-1083.

41.   Boyle, J., et al. *Regression forecasting of patient admission data*. in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. 2008. IEEE.

42.   Combes, C., F. Kadri, and S. Chaabane. *Predicting Hospital Length of Stay Using Regression Models: Application To Emergency Department*. in *10ème Conférence Francophone de Modélisation, Optimisation et Simulation-MOSIM'14*. 2014.

43.   Garcia, K.A. and P.K. Chan. *Estimating Hospital Admissions with a Randomized Regression Approach*. in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*. 2012. IEEE.

44.   Cover, T. and P. Hart, *Nearest neighbor pattern classification.* Information Theory, IEEE Transactions on, 1967. **13**(1): p. 21-27.

45.   Arroyo, J. and Mat\'e, Carlos, *Forecasting histogram time series with k-nearest neighbours methods.* International Journal of Forecasting, 2009. **25**(1): p. 192-207.

46.   Davis, G.A. and N.L. Nihan, *Nonparametric Regression and Short-Term Freeway Traffic Forecasting.* Journal of Transportation Engineering, 1991.

47.   Zhang, L., et al., *An improved k-nearest neighbor model for short-term traffic flow prediction.* Procedia-Social and Behavioral Sciences, 2013. **96**: p. 653-662.

48.   Al-Qahtani, F.H. and S.F. Crone. *Multivariate k-nearest neighbour regression for time series data — A novel algorithm for forecasting UK electricity demand*. in *Neural Networks (IJCNN), The 2013 International Joint Conference on*. 2013.

49.   Tsakoumis, A., S. Vladov, and V. Mladenov. *Daily load forecasting based on previous day load*. in *Neural Network Applications in Electrical Engineering, 2002. NEUREL'02. 2002 6th Seminar on*. 2002. IEEE.

50.   Sapankevych, N.I. and R. Sankar, *Time series prediction using support vector machines: a survey.* Computational Intelligence Magazine, IEEE, 2009. **4**(2): p. 24-38.

51.   Organization, W.H., *International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index.* 1978.

52.   Chatfield, C., *The analysis of time series: an introduction*. 2013: CRC press.
53.   Kane, M.J., et al., *Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks.* BMC bioinformatics, 2014. **15**(1): p. 276.
54.   Box, G.E.P. and G. Jenkins, *Time Series Analysis, Forecasting and Control*. 1990: Holden-Day, Incorporated.
55.   Hyndman, R.J. and Y. Khandakar, *Automatic time series forecasting: the forecast package for R.* Journal of Statistical Software, 2008. **26**(3): p. 1-22.
56.   Team, R.C., *R: A Language and Environment for Statistical Computing*. 2013, R Foundation for Statistical Computing: Vienna, Austria.
57.   Altman, N.S., *An introduction to kernel and nearest-neighbor nonparametric regression.* The American Statistician, 1992. **46**(3): p. 175-185.
58.   Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
59.   Friedman, J., T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Vol. 1. 2001: Springer series in statistics Springer, Berlin.
60.   Vapnik, V., *The nature of statistical learning theory*. 2013: Springer Science & Business Media.
61.   Vert, J.-P., K. Tsuda, and B. Schölkopf, *A primer on kernel methods.* Kernel Methods in Computational Biology, 2004: p. 35-70.
62.   Smola, A.J. and B. Schölkopf, *A tutorial on support vector regression.* Statistics and computing, 2004. **14**(3): p. 199-222.
63.   Cleveland, W.S., E. Grosse, and W.M. Shyu, *Local regression models.* Statistical models in S, 1992: p. 309-376.
64.   Hyndman, R.J. and A.B. Koehler, *Another look at measures of forecast accuracy.* International journal of forecasting, 2006. **22**(4): p. 679-688.
65.   Shcherbakov, M.V., et al., *A survey of forecast error measures.* World Applied Sciences Journal, 2013. **24**: p. 171-176.
66.   Willmott, C.J. and K. Matsuura, *Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance.* Climate research, 2005. **30**(1): p. 79.
67.   Chai, T. and R. Draxler, *Root mean square error (RMSE) or mean absolute error (MAE)?* Geoscientific Model Development Discussions, 2014. **7**: p. 1525-1534.
68.   Hyndman, R.J., *Another look at forecast-accuracy metrics for intermittent demand.* Foresight: The International Journal of Applied Forecasting, 2006. **4**(4): p. 43-46.
69.   Lee, L.H., et al., *Optimizing weekend availability for sophisticated tests and procedures in a large hospital.* Am J Manag Care, 2005. **11**(9): p. 553-8.
70.   Forster, A.J., et al., *The effect of hospital occupancy on emergency department length of stay and patient disposition.* Academic Emergency Medicine, 2003. **10**(2): p. 127-133.