MAPLE TREE
PUBLISHING HOUSE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION (IRJMT)

http://www.mapletreejournals.com/index.php/IRJMT
Received 20 March 2019 ISSN 2582-1040
Accepted 23 March 2019 2019; 1(2); 19-31
Published online 25 March 2019

Visualizing Emergency Department Admission through Data Mining

Shalini Dhiman[1*], Kalaivani S[1], T.K.P. Rajagopal [2]

[1]PG Scholar, Department of Computer science & Engineering, Kathir College Of Engineering, Coimbatore, TN, India

[2]Associate Professor, Department of Computer science & Engineering, Kathir College Of Engineering, Coimbatore, TN, India

**Corresponding author E-Mail ID: shalini1695@gmail.com*

DOI: https://doi.org/10.34256/irjmt1923

ABSTRACT

Emergency Department (ED) boarding –the inability to transfer emergency patients to inpatient beds- is a key factor contributing to ED overcrowding. This paper presents a novel approach to improving hospital operational efficiency and, therefore, to decreasing ED boarding. Using the historic data of 15,000 patients, admission results and patient information are correlated in order to identify important admission predictor factors. For example, the type of radiology exams prescribed by the ED physician is identified as among the most important predictors of admission. Based on these factors, a real-time prediction model is developed which is able to correctly predict the admission result of four out of every five ED patients. The proposed admission model can be used by inpatient units to estimate the likelihood of ED patients' admission, and consequently, the number of incoming patients from ED in the near future. Using similar prediction models, hospitals can evaluate their short-time needs for inpatient care more accurately Emergency Department (ED) boarding – the inability to transfer emergency patients to inpatient beds- is a key factor contributing to ED overcrowding. This paper presents a novel approach to improving hospital operational efficiency and, therefore, to decreasing ED boarding. Using the historic data of 15,000 patients, admission results and patient information are correlated in order to identify important admission predictor factors. For example, the type of radiology exams prescribed by the ED physician is identified as among the most important predictors of admission. The proposed admission model can be used by inpatient units to estimate the likelihood of ED patients' admission, and consequently, the number of incoming patients from ED in the near future. Using similar prediction models, hospitals can evaluate their short-time needs for inpatient care more accurately. We use three algorithms to build the predictive models: (1) logistic regression, (2) decision trees, and Analytic tools (accuracy=80.31%, AUC-ROC=0.859) than the decision tree accuracy=80.06%, AUC-ROC=0.824) and the logistic regression model (accuracy=79.94%, AUC-ROC=0.849). Drawing on logistic regression, we identify several factors related to hospital admissions including hospital site, age, arrival mode, triage category, care group, previous admission in the past month, and previous admission in the past year. From a different perspective, the research focuses on mobility data instead of personal data in general using Structural Equation Modelling analysis method. Based on this research finding, we identified an unexplored factor that can be used to predict the intention to disclose mobility data, and the result also confirmed that context aspects such as demographics and different personal data categories.

Keywords: Data Mining, Emergency Department, Hospitals, Machine Learning, Predictive Models

1. INTRODUCTION

Emergency department (ED) crowding can have serious negative consequences for patients and staff, such as increased wait time, ambulance diversion, reduced staff morale, adverse patient outcomes such as increased mortality, and cancellation of elective procedures. Patients attending the ED typically go through several stages between the time of arrival and discharge depending on decisions made at preceding stages. ED attenders can arrive either via the main reception area or in an ambulance. At this point, the patient's details are recorded on the main ED administration system, before the patient is either admitted, as in severe cases, or proceeds to the waiting area. The patient then waits for a target time of less than fifteen minutes before triage by a specialist nurse. The Manchester Triage scale is used by all Northern Ireland hospitals, and involves prioritizing patients based on the severity of their condition, and to identify patients who are likely to deteriorate if not seen urgently and those who can safely wait to be seen [18]. Triage is an important stage in the patient journey to ensure the best use of resources, patient satisfaction, and safety [19]. Triage systems have also been found to be reliable in predicting admission to hospital, but are most reliable at extreme points of the scale, and less reliable for the majority of patients who fall in the mid points [18].

Once triaged, the patient returns to the waiting room, before assessment by a clinician, who will make a recommendation on the best course of action, which could include treatment, admission, follow up at an outpatient clinic or discharge. If there is a decision to admit the patient, the ED sends a bed request to the ward, and the patient continues to wait until the bed is available. Bottlenecks or excess demand at any point in this process can result in ED overcrowding. Routine recoding of data on hospital administrative systems takes place at each stage of this process, providing an opportunity to use machine learning to predict future stages in the process, and in particular, whether there is an admission.

This study draws on this data to achieve two objectives. The first is to create a model that accurately predicts admission to hospital from the ED department, and the second is to evaluate the performance of common machine learning algorithms in predicting hospital admissions. Previous research has shown ED crowding to be a significant international problem, making it crucial that innovative steps are taken to address the problem here are a range of possible causes of ED crowding depending on the context, with some of the main reasons including increased ED attendances, inappropriate attendances, a lack of alternative treatment options, a lack of inpatient beds, ED staffing shortages, and closure of other local ED departments. The most significant of these causes is the inability to transfer patients to an inpatient bed, making it critical for hospitals to manage patient flow and understand capacity and demand for inpatient beds. One mechanism that could help to reduce ED crowding and improve patient flow is the use of data mining to identify patients at high risk of an inpatient admission, therefore allowing measures to be taken to avoid bottlenecks in the system. Such a model could be developed using data mining techniques, which involves examining and analyzing data to extract useful information and knowledge on which decisions can be taken. This typically involves describing and identifying patterns in data and making predictions based on past patterns. This study focuses on the use of machine learning algorithms to develop models to predict hospital admissions from the emergency department, and the comparison of the performance of different approaches to model development. Patients attending the ED typically go through several stages between the time of arrival and discharge depending on decisions made at preceding stages. ED attenders can arrive either via the main reception area or in an ambulance. At this point, the patient's details are recorded on the main ED administration system, before the patient is either admitted, as in severe cases, or proceeds to the waiting area. The patient then waits for a target time of less than fifteen minutes before triage by a specialist nurse. Triage is an important stage in the patient journey to ensure the best use of resources, patient satisfaction, and safety [19]. Triage systems have also been found to be reliable in predicting admission to hospital, but are most reliable at extreme points of the scale, and less reliable for the majority of patients who fall in the mid points [18].

Once triaged, the patient returns to the waiting room, before assessment by a clinician, who will make a recommendation on the best course of action, which could include treatment, admission, follow up at an outpatient clinic or discharge. If there is a decision to admit the patient, the ED sends a bed request to the ward, and the patient continues to wait until the bed is available. Bottlenecks or excess demand at any point in this process can result in ED overcrowding. Routine recoding of data on hospital administrative systems takes place at each stage of this process, providing an opportunity to use machine learning to predict future stages in the process, and in particular, whether there is an admission.

2. RELATED WORK

Sun et al. [4] developed a logistic regression model using two years of routinely collected administrative data to predict the probability of admission at the point of triage. Risk of admission was related to age, ethnicity, arrival mode, patient acuity score, existing chronic conditions, and prior ED attendances or admission in the past three months. Although their data showed the admission of more females than males, sex was not significant in the final model. Qui et al. [11] used a relative vector machine to predict whether an ED attender would be discharged or admitted to one of three hospital words. Their model had an overall accuracy of 91.9% with an AUC of 0.825. However, the accuracy of predicting the target ward varied by ward and by the probability threshold used.

Lucini et al. [15] used eight common machine learning algorithms to predict admissions from the ED department based on features derived from text recorded on the patient's record. Six out of the eight algorithms had similar levels of performance including nu-support vector machines, support vector classification, extra trees, logistic regress, random forests, and multinomial naive bayes, with AdaBoost and a decision tree performing worst. Taking a different approach, Cameron et al.[17] compared the accuracy of nurses predictions of ED admissions with those of an objective score. They find nurses to be more accurate in cases where they are certain the patient will be admitted, but less accurate than the objective score in cases where they are uncertain about the patient's likelihood of admission. Cameron et al. [2] developed a logistic regression model to predict the probability of admissions at triage, using two years of routine administration data collected from hospitals in Glasgow. The most important predictors in their model included 'triage category, age, National Early Warning Score, arrival by ambulance, referral source, and admission within the last year' (pg. 1), with an area under the curve of the receiver operating characteristic (AUC-ROC) of 0.877. Other variables including weekday, out of hour's attendances, and female gender, were significant but did not have high enough odds ratios to be included in the final models. Kim et al. [21] used routine administrative data to predict emergency admissions, also using a logistic regression model. However, their model was less accurate with an accuracy of 76% for their best model.

The literature highlights the application of a range of traditional and machine learning approaches to the prediction of ED admissions in different contexts using a variety of data. However, there are gaps in the literature to which this study contributes. Much of the previous work focuses on a narrow range of algorithms, and primarily logistic regression, with fewer studies comparing multiple approaches. This leaves open the potential for the development of more accurate predictive models using other algorithms. For example, gradient boosted machines (GBM) were not applied in any of the studies reviewed, but have been successful in predicting binary outcomes in other scenarios such as hospital transfers and mortality [29].

Using a range of clinical and demographic data relating to elderly patients, La Mantiana et al. [9] used logistic regression to predict admissions to hospital, and ED re-attendance. They predicted admissions with moderate accuracy, but were unable to predict ED re-attendance accurately. The most important factors predicting admission were age, Emergency Severity Index (ESI) triage score, heart rate, diastolic blood pressure, and chief complaint [9] (pg. 255). Although these models highlight the usefulness of logistic regression in predicting ED admissions, Xie [22] achieved better performance using a Coxian Phase model over logistic regression model, with the former AUC-ROC of 0.89, and the latter 0.83. Although these models highlight the usefulness of logistic regression in predicting ED admissions, Xie [22] achieved better performance using a Coxian Phase model over logistic regression model, with the former AUC-ROC of 0.89,

Then latter 0.83. Wang et al. [23] used a range of machine learning algorithms to predict admissions from the ED, comparing the ability of fuzzy min-max neural networks (FMM) to other standard data mining algorithms including classification and regression trees (CART), Multi-Layer Perceptron (MLP), random forest, and AdaBoost. Overall, MLP and Random Forest models were the most accurate, both predicting just over 80% of cases correctly, with FMM (with a genetic algorithm) predicting 77.97% of cases correctly. Peck et al. [24] developed three models to predict ED admissions using logistic regression models, naive Bayes, and expert opinion. All three techniques were useful in predicting ED admissions. Variables in the model included age, arrival mode, emergency severity index, designation, primary complaint, and ED provider. Their logistic regression model was the most accurate in predicting ED admissions, with an AUC-ROC of 0.887. Perhaps surprisingly, this model performed better than triage nurse's opinion regarding likely admission. The use of logistic regression to predict admission was subsequently found to be generalizable to other hospitals [10].

Using simulation models, Peck et al. [25] have shown that the use of the predictive models to priorities discharge or treatment of patients can reduce the amount of time the patient spends in the ED department. Baumann and Strout [20] also find an association between the ESI and admission of patients aged over 65. Boyle et al. [2] used historical data to develop forecast models of ED presentations and admissions. Model performance was evaluated using the mean absolute percentage error (MAPE), with the best attendance model achieving a MAPE of around 7%, and the best admission model achieving MAPE of around 2% for monthly admissions. The use of historical data by itself to predict future events has the advantage of allowing forecasts further into the future, but has the disadvantage of not incorporating data captured at arrival and through triage, which may improve the accuracy of short term forecasting of admissions.

3. EXISTING SOLUTION

Many factors contribute to ED boarding. Major increases in hospital admissions and ED presentations with no increase in the capacity of hospitals, a lack of inpatient beds, inadequate or inflexible nurse to patient staffing ratios, inefficient diagnostic services, delays in discharging hospitalized patients, and delays in cleaning rooms after patient discharge have been reported as possible sources of ED boarding (Asplin, 2003; Forero, 2010; Forero, 2011). Additionally, hospital operational inefficiency and lack of communication between inpatient units and ED is a major contributor to ED boarding. Common solutions proposed for ED boarding and crowding are as follows.

- Increasing inpatient capacity
- Altering elective surgical schedules
- Moving admitted ED boarded patients to inpatient hallways ,
- Improving hospital operational efficiency.

No single one of these solutions is always the best option. Increasing hospital capacity can mitigate the problem of overcrowding in most cases, but it is a strategic decision that requires significant time and investment. Altering elective surgical schedules can present a temporary solution that only provides more short-term surgical capacity and does not help patients in need of other critical care (such as ICU). Moving patients to hallways is a controversial solution. While some scholars and ED managers argue in favor of this solution, others believe it may worsen the problem of ED boarding. I believe improving hospital operational efficiency is the key answer to ED boarding. Operational improvement can provide a quick, low-cost, practical solution to ED boarding. For example, Amaras Ingham et al. (2010)'s study shows that an improvement in the admissions protocol in a hospital in Dallas, Texas,. This study explores a scientific approach to improving hospital operational efficiency and, thus, to decreasing ED boarding. The goal is to develop a real-time prediction model capable of estimating the likelihood of admission of each ED patient to the hospital (as inpatient) with a high level of accuracy. These estimations of admission results can be used by inpatient units to estimate the number of incoming patients from the ED. Using4the proposed prediction model, hospitals can more accurately evaluate their short-time needs for inpatient cares. Better estimation of required resources may improve hospital preparedness to provide care for patients arriving from EDs, quicken the process of inpatient bedding, and consequently help reduce ED boarding.

4. PROPOSED SYSTEM

4.1 Methodology

In this study, eight candidate predictor factors were considered for possible inclusion in the model: age, gender, marital status, arrival mode, day and time of ED arrival, encounter reason (chief complaint), and type of radiology exam prescribed by the ED physician (if any). In the interest of analyzing the effect of these factors on the likelihood of the patient's admission to the hospital, the output (target) variable is defined with the two possible values of admission or discharge (rejection).After cleaning the data and transforming it from unprocessed hospital reports to structured records and fields, the analysis was performed in four main steps:

Step1. Descriptive analysis of each predictor factor: each of the eight predictor factors for all the admitted and discharged patients undergoes an exploratory investigation. Two continuous variables corresponding to age and arrival time factors and six categorical variables for the other six predictor factors are defined. Then, using histograms and bar charts, the graphical distribution of each continuous and categorical variable is presented.

Step2. Determining the importance of each predictor factor (variable): each predictor variable is defined and described, after which a "test of significance" is performed. For each continuous variable, an F-test to compare the variable means for the admitted group and discharged group is used; for each categorical variable, a Chi-Square test to compare the frequency of admission in each category of the variable is used.

Step3. Finding relationships between independent variables and target variable in the form of admission rules: In the next step, a C5.0 rule induction algorithm is employed to find relationships between the predictor variables and the output variable, as well as to identify the predictor variables' importance (the C5.0 algorithm is explained in the

Analytical Tools section). Based on the data, a set of rules for the admission of a new patient are discovered. These rules estimate the likelihood of each patient's admission based on his/her predictor variables.

Step4. Developing admission prediction models using independent variables to estimate the target variable: two prediction models based on all eight independent variables are developed, one using the Logistic Regression (LR) technique and the other using Artificial Neural Networks (ANN). The results of these two prediction models are then presented and compared. model

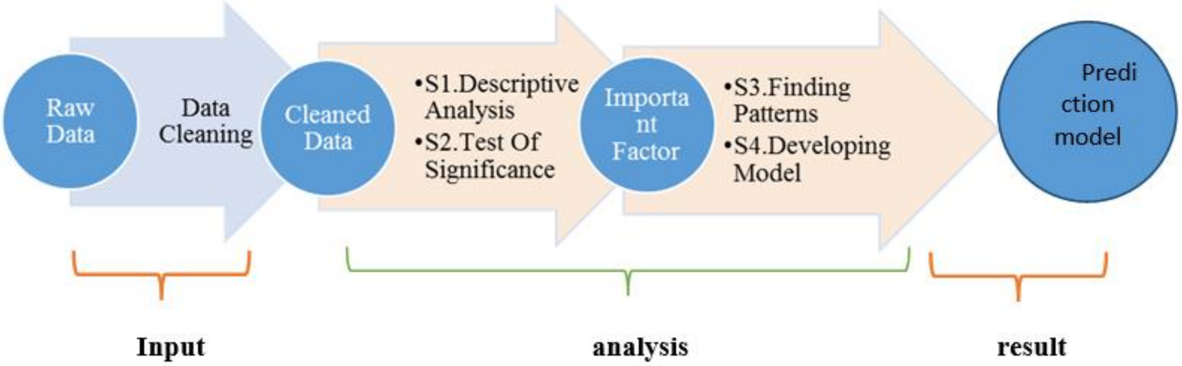

Fig 1.Four main steps of the analysis

4.1.1Raw Data

Raw data (sometimes called source data or atomic data) is data that has not been processed for use. A distinction is sometimes made between data and information to the effect that information is the end product of data processing. Raw data that has undergone processing is sometimes referred to as cooked data. Although raw data has the potential to become "information," it requires selective extraction, organization, and sometimes analysis and formatting for presentation.

4.1.2 Data Cleaning

Real-word data tend to be incomplete, noisy, and inconsistent. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data

4.2 Analytical Tools

Three analytical techniques, namely C5.0 algorithm, Logistic Regression (LR), and Artificial Neural Networks (ANN), are used in this study. The following provides a brief introduction to these three methods.

4.2.1 Classification

Classification is one of the data mining technique which is useful for predicting group membership for data instances. Classification is a supervised kind of machine learning in which there is provision of labeled data in advance. By providing training the data can be trained and we can predict the future of data. Prediction is in the form of predicting the class to which data can belong. Training is based on the training sample provided. Basically there are two types of attributes available that are output or dependent attribute and input or the independent attribute [9]. In the supervised classification, there is mapping of input data set to finite set of discrete class labels. Input data set X € Ri, where i is the input space dimensionally and discrete class label Y €

1......T, where T is the total number of class types. And this is modeled in the term of equation Y=Y(x, w), w is the vector of adjustable parameters.

4.2.2. C5.0 Algorithm

A C5.0 algorithm is a classification technique based on C4.5 by Quinlan (1992). This method can be used to build decision trees and rule sets. A decision tree is a straightforward description of the splits found by the algorithm. In contrast, a rule set is a set of rules that tries to make predictions for individual records. The C5.0 algorithm divides the sample data based on the field that provides the “maximum information gain.” Each division defined by the first split is then divided again and the process repeats until the subsamples cannot be divided further (SPSS Modeler users’ guide, 2012). The C5.0 algorithm is also able to identify predictor variables’ importance in predicting the target variable. The algorithm uses the same criteria (“maximum information gain”) for identifying the importance of predictor variables.

4.2.3. Logistic Regression

Logistic Regression (LR) is a statistical technique for data classification and prediction. In contrast to linear regression, the output variable in Logistic Regression is categorical.LR works by “building a set of equations that relate the predictor variables values to the probabilities associated with each of the output variable categories” (SPSS Modeler users’ guide, 2012). After developing an LR model using available data, it can be used to estimate the value (category) of output variables for new entities. In order to estimate output value, LR calculates the probabilities of membership in every output category and assigns the output value (category) with the highest probability to that entity (Christensen, 1997; SPSS Modeler users’ guide, 2012). Like linear regression, Logistic Regression provides a coefficient value and each predictor variable contribution to variations in the output variable (Menard, 2002).

4.2.4. Artificial Neural Networks

An Artificial Neural Network (ANN) is a mathematical model that attempts to simulate the human brain by collecting and processing data for the purpose of “learning”(Golmohammadi, 2011). An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal from one artificial neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it.

In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs. The connections between artificial neurons are called 'edges'. Artificial neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times. ANNs have different structures and processing algorithms.Figure2 shows a number of well-developed ANN structures.

The feed forward neural network was the first and simplest type. In this network the information moves only from the input layer directly through any hidden layers to the output layer without cycles/loops. Feed forward networks can be constructed with various types of units, such as binary McCulloch-Pitts neurons, the simplest of which is the perceptron. Continuous neurons, frequently with sigmoidal activation, are used in the context of back propagation. Audial basis functions are functions that have a distance criterion with respect to a center. Radial basis functions have been applied as a replacement for the sigmoidal hidden layer transfer characteristic in multi-layer perceptron's. RBF networks have two layers: In the first, input is mapped onto each RBF in the 'hidden' layer. The RBF chosen is usually a Gaussian. In regression problems the output layer is a linear combination of hidden layer values representing mean predicted output. The interpretation of this output layer value is the same as a regression model in statistics. In classification problems the output layer is typically a sigmoid function of a linear combination of hidden layer values, representing a posterior probability. Performance in both cases is often improved by shrinkage techniques, known as ridge regression in classical statistics. This corresponds to a prior belief in small parameter values (and therefore smooth output functions) in a Bayesian framework.

RBF networks have the advantage of avoiding local minima in the same way as multi-layer perceptron's. This is because the only parameters that are adjusted in the learning process are the linear mapping from hidden layer to output layer. Linearity ensures that the error surface is quadratic and therefore has a single easily found minimum. A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit temporal dynamic behavior for a time sequence. Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition.

The term "recurrent neural network" is used indiscriminately to refer to two broad classes of networks with a similar general structure, where one is finite impulse and the other is infinite impulse. Both classes of networks exhibit temporal dynamic behavior.[4] A finite impulse recurrent network is a directed acyclic graph that can be unrolled and replaced with a strictly feedforward neural network, while an infinite impulse recurrent network is a directed cyclic graph that cannot be unrolled.

Both finite impulse and infinite impulse recurrent networks can have additional stored state, and the storage can be under direct control by the neural network. The storage can also be replaced by another network or graph, if that incorporates time delays or has feedback loops. Such controlled states are referred to as gated state or gated memory, and are part of long short-term memory's (LSTMs) and gated recurrent units. In the field of mathematical modeling, a radial basis function network is a network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks have many uses, including function approximation, time series prediction, classification, and system control. They were first formulated in a 1988 paper by Broom head and Lowe, both researchers at the Royal Signals and Radar Establishment. Radial basis function (RBF) networks typically have three layers: an input layer, a hidden layer with a non-linear RBF activation function and a linear output layer.

This study uses a Multiplayer Perceptron (MLP), one of the most common forms of ANNs. A multilayer perceptron (MLP) is a class of feedforward artificial neural network. An MLP consists of, at least, three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

Multilayer perceptron's are sometimes colloquially referred to as "vanilla" neural networks, especially when they have a single hidden layer each continuous variable, an F-test to compare the variable means for the admitted group and discharged group is used; for each categorical variable, a Chi-Square test to compare the frequency of admission in each category of the variable is used.

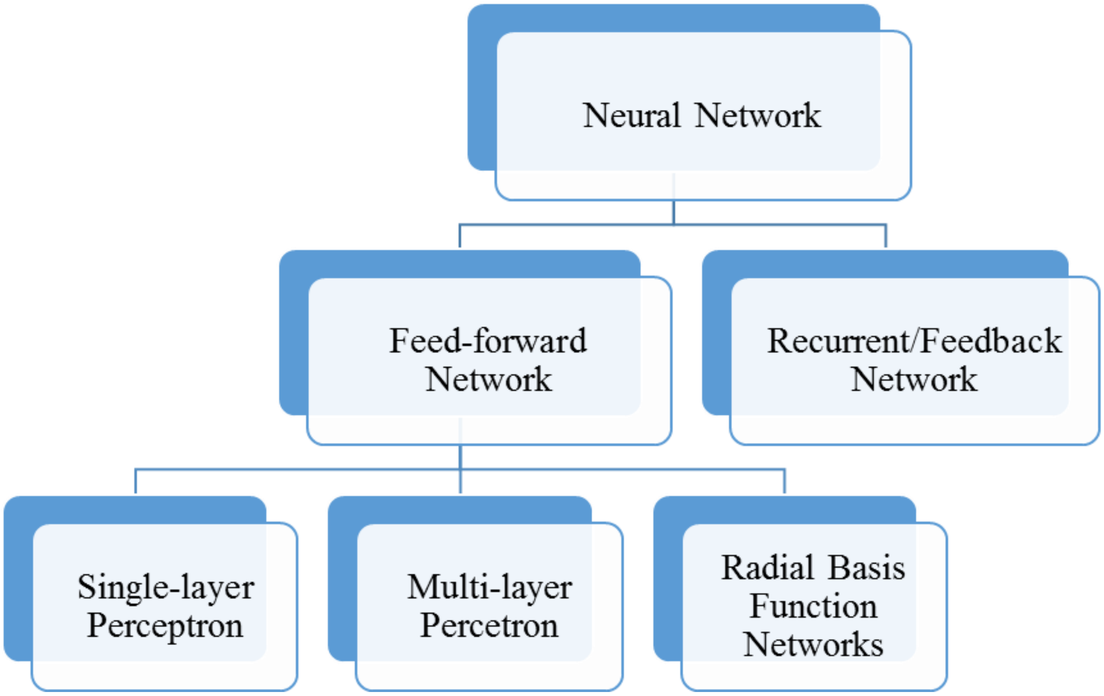

Fig 2. A taxonomy of Neural Network architectures (after Gardner and Dorling, 1998)

Unlike many statistical techniques, the MLP makes no assumptions on the distribution of data, the linearity of the output function, or the type (measurement) of predictor and output variables (Gardner and Dorling, 1998; SPSS Modeler users' guide, 2012). An MLP consists of multiple parallel layers of nodes, which are connected by weighted links as shown in Figure3. The input layer contains the independent variables, the middle layers (hidden layers) contain the processing units, and the output layer contains the output variable(s). The process of finding the right weights in an ANN is called training. Training consists of two general phases of assigning weights and updating them to minimize the model's error (Golmohammadi et al., 2009; Golmohammadi, 2011). These phases are repeated until the performance of the network is satisfactory. In an MLP, the weights are usually estimated using Back propagation (backward propagation of errors), a generalization of the Least Mean Squares algorithm (Gardner and Dorling, 1998).

4.3 Prediction models

The, the performances of these prediction models on the historic data were calculated and compared. Before developing the models, some modification to data were required. The major modification was related to missing information for some observations. After eliminating the observations with missing data, the total number of 10380 visits remained as input data for the LR prediction model

4.4 LR Prediction Model

Its basic form, uses a logistic function to model a binary dependent variable; many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model; it is a form of binomial regression. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick; these are represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds for the value labeled "1" is a combination of one or more independent variables the independent variables can each be a binary variable or a continuous variable . The corresponding probability of the value labeled "1" can vary between 0 and 1 Using SPSS Modeler (V15.0)'s Logistic Regression tool, an LR model with Binominal output was developed (since the target variable, admission result, has only two possible values). Three common LR methods, "Enter," "Forwards," and "Backwards," were tested and the highest level of accuracy was obtained using the "Enter" method.

Two of the predictor categorical variables, encounter reason and radiology exam, include almost 200 categories each. Therefore, the generated LR function (to estimate the target) is extremely large. However, the Modeler software enabled us to perform a sensitivity analysis of the LR model and to calculate the weights assigned to each predictor variable. These weights show the effect of each predictor variable in estimating the target variable and can be translated as the predictor variable's importance in predicting the target variable (admission result).

4.5 ANN Prediction Model

Advantage of ANN to develop the second prediction model. In developing an ANN, the number of hidden layers (or nodes) and initial weights need to be set. In addition, I needed to decide what portion of data to use for training, choose a learning algorithm, and define a stopping rule for the training procedure. Using SPSS Modeler (V15.0)'s ANN method, several different structures with different numbers of hidden nodes (in one and two hidden layers) were tried. The results, then, were compared to the SPSS Modeler's recommended ANN structure. The highest level of accuracy for ANNs developed based on the predictor variables and available data was achieved with a model with 14 hidden nodes in one layer.

5. MODEL PERFORMANCE

Table 1: Model Performance

	Accuracy (%)	**Kappa**	**AUC-ROC**	**Specificity**	**Sensitivity**
Logistic Regression	79.94	0.4600	0.8497	0.8995	0.5357
Decision Tree	80.06	0.4661	0.8249	0.9015	0.5349
(RPART) GBM	80.31	0.4724	0.859	0.9038	0.5379

We used accuracy, kappa, AUC-ROC, sensitivity and specificity to evaluate the predictive performance of the models by making predictions on the test data. As shown in table 2, the GBM performs best across all performance measures. However, in some cases differences in performance across the models are small. Logistic regression and decision tree models show similar levels of predictive performance, with the decision tree performing only slightly better than the logistic regression model in terms of accuracy and kappa, and the logistic regression model performing better in terms of AUC-ROC and sensitivity. As a consequence of the class imbalance, specificity is considerably higher than sensitivity across all three models. These findings corroborate with those of Lucini et al. [27] who report similar levels of performance across the majority of models presented in their study.

6. DISCUSSION

Using the available data of patients, I was able to discover patterns between patients' characteristics, identify the important factors in patients' admission to hospital, and develop an admission prediction model. Here, I further discuss two issues related to the model input and output, one a conceptual issue about the relationship between the input and the output, and the other, a technical issue about the output. The first issue arises from the difference between causal and correlation relationship between predictor factors and the result. The discovered patterns and developed models in this study are all based on the correlation relationships between the predictor factors and the admission results. Although some factors, such as encounter reason, may have a causal effect on the admission result, the predictor factors discovered in this study should be considered as co relational factors. The purpose of the models in this study is to serve 40 as a real time predictor of the admission results for new patients, not to find the causes of their admissions. The second issue is related to destinations of the patients. Given the limitation of the available data, the result of the developed models is patients' admissions or discharges. Although this information provides great insight for the ED and hospital, it only can drive an estimation of the total demand for all inpatient units. This information can be communicated to all inpatient units, such as ICU and operating rooms, as an estimation of their combined demand, but it cannot determine the demand for each unit. I acknowledge that having the demand for each unit can contribute to the decrease in ED boarding and ED overcrowding more than the combined demand, in most cases. This study provides a foundation for developing extended models with more detailed outputs, when the required data is available. This study suggests that in order to decrease ED overcrowding and boarding, hospital and ED managers should focus more on operational efficiency and communication. I believe hospital units, including ED, need to become more "connected". Instead of focusing on each unit's output, managers need to see hospital as a whole system and focus on increasing the system's output. By estimating the real time inpatients demands (from ED) and communicating them to inpatient units, the proposed prediction models provide unit managers with an extra piece of information about their units' demands. Managers can incorporate this information in 41 their real time decision makings process, and over time, they will be able to make more informed and accurate decisions about their resource utilization and allocation. The implementation of this study in an ED requires an integrated information sharing system, for communicating the estimates of demands, from the ED to inpatient units. In addition, a user interface for inputting new patients' information into the system and a simple processor machine (or a desktop computer) for running the model in required.

7. CONCLUSION

The main purpose of this study was to find an effective and efficient operational solution to the problem of patient boarding in emergency departments. One of the main causes of ED boarding is that inpatient units do not have an accurate and timely estimation of the number of near-future incoming ED patients. I tried to find a solution to estimate the number of ED patients in need of inpatient cares earlier and more accurately. This goal was achieved by developing real-time admission prediction models capable of estimating the likelihood of admission for each ED patient using the patient's information. These estimations then can be used by inpatient units to create a better estimate of their incoming patients in near-future. Based on the historic data of 15,000 ED patients from a local hospital in the Boston area, eight important predictor factors of the admission result were identified: patient age, arrival time at ED, marital status, gender, arrival mode, day of arrival, encounter reason, and radiology test prescribed by the ED physician. After exploring each of these factors, age, encounter reason, and radiology exams were identified as the most important predictor factors of patients' admission to the hospital. To the best of my knowledge, this research is the first work to study the effect of different types of radiology exams prescribed by the ED physician on the patients' admission results.

REFERENCES

[1] J.S. Olshaker, N.K. Rathlev, "Emergency Department overcrowding and ambulance diversion: The impact and potential solutions of extended boarding of admitted patients in the Emergency Department", J. Emerg. Med. 30 (2006) 351–356. doi:10.1016/j.jemermed.2005.05.023.

[2] A. Cameron, K. Rodgers, A. Ireland, R. Jamdar, G.A. McKay, A simple tool to predict admission at the time of triage., Emerg. Med. J. 32 (2013) 174–9. Doi: 10.1136/emermed-2013-203200.

[3] J. Boyle, M. Jessup, J. Crilly, D. Green, J. Lind, M. Wallis, P. Miller, G. Fitzgerald, Predicting emergency department admissions, Emerg. Med. J. 29 (2012) 358–365. doi:10.1136/emj.2010.103531.

[4] S.L. Bernstein, D. Aronsky, R. Duseja, S. Epstein, D. Handel, U. Hwang, M. McCarthy, K.J. McConnell, J.M. Pines, N. Rathlev, R. Schafermeyer, F. Zwemer.

[5] Y. Sun, B.H. Heng, S.Y. Tay, E. Seow, Predicting hospital admissions at emergency department triage using routine administrative data, Acad. Emerg. Med. 18 (2011) 844–850. doi:10.1111/j.1553-2712.2011.01125.x.

[6] Amarasingham, R. et al., 2010. A rapid admission protocol to reduce emergency department boarding times. *Quality safety in health care*, 19(3), pp.200–204

[7] Hoot, N.R. & Aronsky, D., 2008. Systematic review of emergency department crowding: causes, effects, and solutions. Annals of emergency medicine, 52(2), pp.126–136.

[8] IBM SPSS Modeler. IBM. Available at: http://pic.dhe.ibm.com/infocenter/spssmodl/-v15r0m0.

[9] S.W. Kim, J.Y. Li, P. Hakendorf, D.J.O. Teubner, D.I. Ben-Tovim, C.H. Thompson, Predicting admission of patients by their presentation to the emergency department, EMA - Emerg. Med. Australas. 26 (2014) 361–367. doi:10.1111/1742-6723.12252.

[10] Ruger, J.P., Lewis, L.M. & Richter, C.J., 2007. Identifying high-risk patients for triage and resource allocation in the ED. The American journal of emergency medicine, 25(7), pp.794–798.

[11] Sadeghi, S. et al., 2006. A Bayesian model for triage decision support. International Journal of Medical Informatics, 75(5), pp.403–411.

[12] S. Qiu, R.B. Chinnam, A. Murat, B. Batarse, H. Neemuchwala, W. Jordan, A cost sensitive inpatient bed reservation approach to reduce emergency department boarding times, Health Care Manag. Sci. 18 (2015) 67–85. Doi: 10.1007/s10729-014-9283-1.

[13] Steele, R. et al., 2006. Clinical decision rules for secondary trauma triage: predictors of emergency operative management. Annals of Emergency Medicine, 47(2), p.135.

[14] F.R. Lucini, F.S. Fogliatto, G.J.C. da Silveira, J. Neyeloff, M.J. Anzanello, R. de S. Kuchenbecker, B.D. Schaan, Text mining approach to predict hospital admissions using early

medical records from the emergency department, Int. J. Med. Inform. 100 (2017) 1–8. doi:10.1016/j.ijmedinf.2017.01.001.

[15] Viccellio, A. et al., 2009. The association between transfer of emergency department boarders to inpatient hallways and mortality: a 4-year experience. Annals of emergency medicine, 54(4), pp.487–491.

[16] A. Cameron, A.J. Ireland, G.A. McKay, A. Stark, D.J. Lowe, Predicting admission at triage: are nurses better than a simple objective score?, Emerg. Med. J. 34 (2017) 2–7. doi:10.1136/emermed-2014-204455.

[17] S. Xie, J. and Coggeshall, Prediction of transfers to tertiary care and hospital mortality: A gradient boosting decision tree approach, Stat. Anal. Data Min. ASA Data Sci. J. 3 (2010) 253–258.

[18] J. Boyle, M. Jessup, J. Crilly, D. Green, J. Lind, M. Wallis, P. Miller, G. Fitzgerald, Predicting emergency department admissions, Emerg. Med. J. 29 (2012) 358–365. doi:10.1136/emj.2010.103531.

[19] J.S. Peck, J.C. Benneyan, D.J. Nightingale, S.A. Gaehde, Characterizing the value of predictive analytics in facilitating hospital patient flow, IIE Trans. Healthc. Syst. Eng. 4 (2014) 135–143. Doi: http://dx.doi.org/10.1080/19488300.2014.930765

Conflict of Interest

None of the authors have any conflicts of interest to declare.

About the License