Acknowledgement

The realization of this work has been through the wonderful guidance, and mentor-ship of my supervisors; Mrs. MARQUET Kristel, Dr. SCHROOTEN Ward, and Prof. dr. Paul Janssen. I am very grateful to them since this work would not have been fruitful without their professional advice.

I convey my gratitude to VLIR-UOS scholarship for awarding me this important opportunity to develop my knowledge and expertise in the field of statistics. I thank all the professors of the master of statistics programme for their guidance and teaching expertise which has really built in me skills to handle different statistical scenarios. I also thank all my classmates and friends not mentioned for making my stay in Belgium lively and enjoyable.

I also wish to extend my gratitude to my family. I am grateful to my mum, Sarah Mwangi, and dad, Patrick Mwangi, for providing their unwavering support over the years and my siblings who have also played important roles in various ways.

Most of all, I thank the almighty God for giving me wisdom, strength and good health to successfully complete the program.

James Wanja
Diepenbeek, Belgium, September 2015

Abstract

Background: One of the challenges in safety improvement of healthcare is the measurement of Adverse events. Retrospective medical record review is one of the methods commonly used to measure the prevalence of adverse events but despite their positive impact, they are faced by some challenges in that they are labor intensive, time-consuming, expensive, and require extensive use of professional review.

Objective: The objective of the study was to predict the presence of an adverse event among patients who had an unplanned transfer to a higher level of care whereby this prediction could be used for future adverse event detection since record review is a very labor intensive procedure.

Data: A retrospective analysis was performed on 830 patients records who had an unplanned transfer to a higher level of care which includes transfer to intensive care or an in-hospital medical emergency team intervention with an objective of evaluating how often such unplanned transfer was associated with an adverse event. From the reviewed patient records, 465 (56%) patients had one or more adverse events and 365 (44%) did not have an adverse event.

Methodology: To improve on some of the shortcoming of the record review, modeling approaches were used whereby the performance of different classification and prediction models were compared in terms of the accuracy in classification. This involved using supervised learning methods, which included regression based methods (e.g. generalized linear models, and penalized regression) and tree based methods (e.g. decision trees and random forests).

Results and conclusion: The performance of the different classifiers were assessed by sensitivity, specificity, the area under the ROC curve (AUROC). AUROC ranged from a low of 66.65% for the classification tree to 68.98% for the stepwise logistic regression which showed the greatest predictive accuracy for the probability of an adverse event. Sensitivity ranged from 46.93% for the random forest to a high of 62.01% for the classification tree. Specificity ranged from 64.05% for the classification tree to a high of 82.35% for the random forest. Stepwise logistic regression had a better specificity-sensitivity (77.93%-53.78%) trade off and the highest (64.34%) correct classification hence it was selected since it performed better compared to the other classifiers.

Key words: Adverse events; Logistic regression; LASSO; Classification tree; Random forests; Cross-validation; ROC.

Contents

List of Tables

List of Figures

1 Introduction

One of the challenges in safety improvement of healthcare is the measurement of adverse events (AEs). AEs are defined as an unintended injury or complication, which results in disability at discharge, death, or prolongation of hospital stay, and is caused by healthcare management (including omissions) rather than the patient's disease (Wilson et al., 1995).

Adverse events associated with medical treatments are a major source of morbidity and mortalit. A study, in Swedish hospitals, estimated that 12.3% of the admission had adverse events (95% CI, 10.8-13.7) of which 70% were preventable. 55% of preventable AEs led to impairment or disability, 9% led to permanent disability and 3% of the adverse events contributed to patient death (Soop et al., 2009). de Vries et al. through a record review study estimated that AEs affect 9.2% of the patients during hospital admission, of which almost half were assessed as preventable. This confirms that adverse events can cause extensive human suffering hence they are an important concern in healthcare.

A patient with an AE may require an unplanned intensive care admission (UIA), which is defined as all patients unexpectedly admitted to the ICU from a lower level of care in the hospital (Baker et al., 2004), and it has been reported that these constitute 1-9% of all ICU admissions (Posa et al., 1992). Hence focusing on unplanned transfers to a higher level of care as a trigger to detect AEs would be more efficient compared to the review of records randomly (Marquet et al., 2015) so in this study special focus was on higher level of care patients.

Health care organizations use a wide array of methods to uncover and monitor AEs which include;- review of medical records, studies based on interviews with health care providers, direct observation, institutional and national reporting systems, analysis of existing and routinely collected data (studies of claims and complaints, information technology and electronic medical records, administrative data, autopsy reports, and mortality and morbidity conferences) and many others (Michel, 2003).

Retrospective medical record review, which is the use of prerecorded patient-focused data as the primary source of information to answer a research question (Worster and Haines, 2004), is one of the methods commonly employed to measure the prevalence of adverse events which is effective in estimating the nature, frequency and economic impact of adverse events.

Despite its positive impact, medical record reviews are faced by some challenges in that they are time-consuming, expensive, and they require extensive use of professional review. The reliability of the record review data also has the limitation of incomplete documentation in some of the medical records which makes their accuracy questionable. Record reviews' reproducibility, that is the property of a measurement tool to yield similar results when repeatedly applied to the same phenomenon, is also a concern. Further, when it comes to cases of international comparisons they need to be interpreted carefully and extra effort has to be put in standardizing the methods across different definitions (Michel, 2003).

To improve on some of the shortcoming of the record review, for example being expensive, time consuming and labor intensive, in this study we propose the use of statistical modeling techniques of prediction and classification. Patients with a need for higher level of care there are of two distinct groups, that is those with an adverse event and those without and the capability to distinguishing with some degree of accuracy between these two groups is particularly important in that it will help in flagging patients records with a particular problem or issue. This will help enable us identify the subset of patients that require/ doesn't require their records reviewed.

Getting the subset of patients characteristics that can best assist in predicting the probability of an adverse event will aid in future adverse event detection and will also guarantee reproducibility when it comes to their application on similar phenomenon in different settings.

Therefore the objective of our analysis is to predict the presence of an adverse event among patients who had an unplanned transfer to a higher level of care, using the data obtained in the patient record review study. To achieve this objective, we compare the performance of different methods in predicting the presence or absence of adverse events and also the accuracy of these classification methods in classifying patients with a need for higher level of care into;- those with an adverse event detected and those without an adverse event.

The remaining part of this report is organized as follows: Section 2 gives a description of the study data used for the analysis, Section 3 describes the statistical methodology used, Section 4 presents the results of the analysis and finally discussion and concluding remarks are presented in Section 5. Some relevant references follow thereafter.

2 Data

A retrospective analysis was performed on 830 patients records who had an unplanned transfer to a higher level of care with an aim of predicting the presence of an adverse event (Marquet et al., 2015). From the reviewed patient records, 465 (56%) patients had one or more adverse events detected and 365 (44%) did not have an adverse event.

During a six month period (between November 7, 2011 and May 6, 2012) the records of all patients with an unplanned need for a higher level of care were reviewed. The review was done in six hospitals in the province of Limburg, Belgium, by the same experienced and independent clinical team consisting of a research nurse (specialized in intensive care, emergency care, and health care management), a physician (specialized in anesthesiology and emergency medicine), and a clinical pharmacist.

The selection of unplanned transfers involved selecting all medical emergency team (MET) interventions and unplanned intensive care admission (UIAs) with exclusion of;- planned ICU admissions, ICU admissions directly from the emergency department, and neonatal and maternal ICUs due to specific nature (Marquet et al., 2015).

As candidate variables for predicting the presence of adverse events or for classifying the patients into either the adverse events group or no adverse events group, we considered patients characteristics collected from the moment of admission to the point where the patient was transferred to a higher level of care. Some of these include patient's sex, age category, type of admission, whether or not they had previous admission within the last 3 months, number of prescribed medication before admission, severity score at admission and many others. All these variables including demographic information are listed in Tables 3, 4, and 5 in the Appendix.

The response of interest (Adverse Events, AE) was dichotomized as follows;

$$AE_i = \begin{cases} 1, & \text{Adverse Event detected} \\ 0, & \text{No Adverse Event detected} \end{cases}$$

3 Methodology

The performance of different classification and prediction methods were compared and assessed. This involved using supervised learning methods, which include regression based methods (like;- generalized linear models - logistic regression, and penalized regression) and tree based methods (including;- decision trees and random forests). The two primary objectives of supervised learning are variable selection and prediction. For this study, variable selection involves choosing a subset of clinical characteristics that are most associated with the presence or absence of an adverse event following an unplanned transfer to higher level of care. Prediction involved using these selected clinical characteristics to develop models that will aid in accurately predicting the probability of having an adverse event within the data (train set) used in the learning process and perform also well on an independent test data (Dasgupta et al., 2011).

Before each model building and with the view to assess the generalization ability of the models (classifiers) used throughout our analysis, the data set was randomly split into two, a training set to aid in the learning bit which was 60% and 40% test set for the validation of the models. This would help assess the most accurate model for predicting an independent test data from the same population.

3.1 Logistic Regression

Logistic regression is useful for situations in which you want to be able to predict the presence or absence of a characteristic or outcome based on values of set of independent variables which are continuous, categorical, or both (Hosmer and Lemeshow, 2000). It is similar to linear regression but suited for models where the response of interest is binary like in our case.

Logistic regression was used to model the relationship between the presence or absence of an adverse event and the set of independent patient characteristics. For p independent variables, the model can be written as in Equation (1)

$$P(Y_i = 1|X_i) = \frac{1}{1 + \exp -(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}, \quad i = 1, \ldots, 830 \tag{1}$$

where;
Y_i is the value of the dichotomous outcome for patient i; 0- No Adverse event, 1- Adverse event detected, $P(Y_i = 1|X_i)$ is the probability of the presence of an adverse event in patient i given the patients characteristics. And X_i is the value of the independent variable patient characteristics for the i^{th} patient while $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients

The linear version of the logistic model is as in Equation (2), this has many desirable properties of a linear regression model where the parameters can be fit without boundary constraints (Hastie et al., 2015).

$$logit[P(Y = 1|X)] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{2}$$

Maximum likelihood is used in estimating the unknown parameters, $\boldsymbol{\beta}$. The principle of maximum likelihood states that we use the values that maximize expression in equation (3) as the values for the $\boldsymbol{\beta}$ (Hosmer and Lemeshow, 2000). So we seek estimates for $\boldsymbol{\beta}$ such that the predicted probability $\hat{P}(Y_i = 1|X_i)$ of an adverse event for each patient, using (1), corresponds as closely as possible to the patient's observed adverse event status.

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}, \qquad i = 1, \ldots, 830 \tag{3}$$

A common problem in statistical modeling is variable selection, that is, which input variables should be retained in a model. Prediction of the outcome was the main objective of our analysis so model with "best" predictions was of interest. Methodologically we mostly choose the model with the smallest deviance (-2 × log-likelihood), but deviance is related to the training error which can be a poor estimate of the test error which we aim to minimize. In order to select the best model with respect to low test error, we need to estimate this test error. To achieve this we used these two common approaches (James et al., 2013);

- Indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting. This was achieved by using stepwise selection where a subset of the p predictors that are associated with the outcome are identified and the model is fit on these reduced set. Adjustment to the train error is then done by applying Mallow's C_p, AIC, or BIC to select a single best model among a set of models.

- Directly estimate the test error, using either a validation set approach or a cross-validation approach. We applied this through the shrinkage/regularization approach. This was achieved by fitting a model involving all the p predictors and shrinking some of the coefficients towards zero using a shrinkage parameter which is chosen through the assessment of the cross-validation error.

3.1.1 Logistic Regression - Stepwise approach

In general, there are 2^p models that involve subsets of p predictors. Best subset selection may suffer from statistical problems when p is large since the larger the search space, the higher the chance of finding models that look good on the training data, but performs poorly on test data. Thus a large search space can lead to over-fitting and high variance of the coefficient estimates (James et al., 2013). Hence, stepwise approach which explores a more restricted set of models was a good alternatives for our study where we had many predictors.

Stepwise model selection approach of either deletion or selection of variables from a model is based on a statistical algorithm that checks the "importance" of variables and either drops or retains them. The importance is defined in terms of the statistical significance of the variable's coefficient and this significance can be assessed by using; Likelihood ratio test, score, or Wald test. At any step the most important variable is one that produces the greatest change in the log likelihood relative to a model not containing the variable (Hosmer and Lemeshow, 2000). However, after adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit.

In this study, selection of terms for deletion or inclusion was based on Akaike's information criterion (AIC). This was performed using `R step` function which uses AIC and the procedure stops when the AIC criterion cannot be improved. Overall fit was assessed by chi-square statistics proposed by Hosmer-Lemeshow goodness of fit test. The function defines AIC follows;

$$\text{AIC} = -2 \text{ maximized log-likelihood} + 2 \text{ number of parameters}$$

To get the variability in the prediction performance of the logistic model on the test set, the model was refit 100 times on different samples of the train data and test data then averaged to get the overall assessment of the predictions on the test data.

3.1.2 Lasso penalty Approach

We can fit a model containing all p predictors using a technique that constrains or that shrinks the sizes of the coefficient estimates towards zero. Two common techniques for shrinking the regression coefficients are ridge regression and the least absolute shrinkage and selection operator (LASSO). In the case of lasso unlike ridge, the penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large. Hence, much like best subset selection, the lasso performs variable selection (James et al., 2013).

The Logistic model was then fit based on maximizing the likelihood, or equivalently minimizing the negative log-likelihood along with an ℓ_1 norm (sum of the absolute values) penalty.

$$minimum_{\beta_0,\beta_j} \left\{ -\frac{1}{N}\mathcal{L}(\beta_0, \beta_j; \mathbf{y}, \mathbf{X}) + \lambda ||\beta_j||_1 \quad j = 1, \ldots, p \right\} \tag{4}$$

where $\mathbf{y}$ is the N-vector of outcomes and $\mathbf{X}$ is the $N \times p$ matrix of predictors, $\mathcal{L}$ is the log-likelihood, and λ is a tuning parameter that determines the degree of shrinkage which controls the strength of the lasso penalty whereby the larger λ the more shrinkage, hence it determines the number of variables forced to zero (Hastie et al., 2015).

Given the logistic model (1), the negative log likelihood with lasso ℓ_1 regularization takes the form;

$$-\frac{1}{N}\sum_{i=1}^{N}\{y_i(\beta_0 + \boldsymbol{\beta}^T x_i) - log(1 + e^{\beta_0 + \boldsymbol{\beta}^T x_i})\} + \lambda ||\beta||_1 \tag{5}$$

Our model was fit using the lasso shrinkage approach. This was fit on the training set and validation done on the test set. To ensure our covariates maintained their group structure, i.e all the coefficients within a group to be shrunk simultaneously, an extension to the lasso, group lasso, was used to select all the coefficients for a particular variable to be in or out of the model (Hastie et al., 2015). The group lasso penalty takes the form, $(\lambda \sum_{g=1}^{G} ||\beta_g||_2)$, where $||\beta_g||_2$ is the Euclidean norm (not squared) for the coefficients of inputs in the g^{th} group (Dasgupta et al., 2011).

Implementing the lasso requires choosing the optimal value of the tuning parameter, λ. This value was chosen from a grid of λ values using cross-validation where the value that results to the smallest cross-validation error (or largest predictive ability) was selected. Finally the model was re-fit using all the available observation and the selected value of the tuning parameter, λ.

To get the variability in the prediction performance of the lasso logistic model on the test set, the model was refit 100 times on different random samples of the train data and test data then averaged to get the overall assessment generalization ability in terms of its predictions on the test data.

3.2 Choosing the Tuning parameters

k-fold cross-validation was employed in choosing the tuning parameters λ for the lasso and also for the tree based approaches. This involves splitting the data into k random subsets, then the algorithm is trained on the $k - 1$ subset and tested on the remaining subset (the test set). This is iterated over all the k possible arrangements of the subsets into the two groups. At each iteration the measure of model performance, for example the prediction error, on the test set is computed leading to k estimates which are then averaged to get the final estimate of the performance. The idea is that the algorithm is trained and tested on statistically independent subsets (Dasgupta et al., 2011).

The value of k is often selected between $5 - 10$, because less than 5 tends to increase the variance of the estimate while more than 10 does not improve the estimation in terms of variance or bias. For our study $k = 10$ was the value used.

The performance measure from the k-fold cross-validation was used in selection the tuning parameters. For the lasso logistic regression, λ was selected from a set of independent variables/predictors that gave the smallest cross-validation error which were also chosen for the final model. For the classification tree, cross validation was also used to determine the number of branches to keep (pruning parameter) from the tree through the minimization of the cross-validation error (Dasgupta et al., 2011).

3.3 Tree Based

3.3.1 Decision Trees - Classification Trees

Decision trees are broadly classified into classification trees, this is where the outcomes are categorical which applies to our case, and regression trees where the outcome is continuous (Breiman et al., 1984). This methodology has the ability to efficiently split populations into meaningful subgroups. It does this by identifying mutually exclusive and exhaustive subgroups of a population whose members share common characteristics that influence the outcome variable of interest. They eventually produce a visual output that is a multilevel structure resembling branches of a tree.

The classification tree procedure examines all possible independent, or splitting, variables and selects the one that results in binary groups that are most different with respect to the outcome variable, according to a predetermined splitting criterion. The parent node then branches into two descendant, or child, nodes according to the independent variable that was selected.

The best predictor (splitting variable) is examined to find the split producing the largest improvement in goodness-of-fit, chosen using an impurity or diversity measures (Gini, entropy, and minimum error). The goal is to produce subsets of the data which are as homogeneous as possible with respect to the outcome variable (Breiman et al., 1984)

In our study, for the impurity measure we used the gini impurity which is used for categorical target variables. The gini diversity index at node t, $g(t)$, is defined as;

$$g(t) = 1 - \sum_{j} p^2(j|t) \tag{6}$$

where; j is the category of the target variable. When cases at a node are evenly distributed between the groups, the gini index taxes a maximum value of 1-(1/number of groups) and when all cases in a node belong to the same category the index equals 0 (Kurt et al., 2008). The variable whose split provides the largest value of gini improvement measure is selected for splitting at each step.

Improvement measure = diversity index of parent node − weighted diversity index of the two child nodes.

Decision trees need to be optimized (pruned) with respect to the number of variables used and the size of the tree grown since a bushy (large) tree may be hard to interpret and may also have too much variance. k-fold cross-validation, with $k = 10$, was used in the optimization by looking for the pruning parameter that minimizes the misclassification error.

Once a tree is learned using training data, the predicted outcome for each observation is determined by sending the input values down the tree and taking the most frequent (vote) class of the test data in the terminal node into which the observation falls.

Classification trees have the advantage that they are typically easy to interpret, and are also non parametric hence no assumptions on the predictor variables are made so it can handle highly skewed, multi-modal data and are robust to outliers (Breiman et al., 1984). However, they tend to have a high variability whereby a slight change in the data can have significant change on the splitting rules that make up the tree (James et al., 2013). A further drawback lies in the fact that continuous variables are implicitly discretized by the splitting process, losing information along the way.

3.3.2 Random Forests

Classification trees tend to overfit the training data and they do not always have good performance on the test set in terms of accuracy. An extension to the classification trees is the random forest and bagging approach which averages over several trees grown on a re-sampled versions of the training data. This averaging can help reduce the variability experience by classification trees (James et al., 2013).

Random forests build many trees, without pruning, on bootstrapped training data. During the building, a random sample, $mtry$, rather than all of the p predictors is chosen each time a split in a tree is considered. The split is allowed to use only one of the *mtry* predictors and a fresh sample of *mtry* predictors is taken at each split. This aims at "decorrelating" the trees thereby making the average of the resulting trees less variable hence increasing the prediction accuracy (James et al., 2013).

There are two important tuning parameters, the number of trees (*ntree*) to be grown and the number of predictors (*mtry*) chosen at each split in growing the tree. For classification, the default value for *mtry* is $\sqrt{p}$ and the minimum node size is one (Dasgupta et al., 2011).

In classification and prediction, the predicted outcome of each observation is estimated by the most frequent predicted outcome from each component tree. That is, for a given test observation, we record the class predicted by each of the bootstrapped (B) trees, and take a majority vote. The overall prediction is the most commonly occurring majority class among the B predictions.

3.3.3 Variable importance in Random Forests

In many situations, the aim is not only to make the most accurate predictions of the response but also to identify which predictor variables are the most important to make these predictions. Random forest also provide variable importance measures which include; mean decrease in accuracy and mean decrease in gini.

Random forest uses the out-of-bag (the remaining observations not used to fit a given tree from the bootstrap samples) samples to construct mean decrease in accuracy measure. This measures the prediction strength of each variable. When a tree is grown the out-of-bag (OOB) sample is passed down the tree and the prediction accuracy recorded. The values of j^{th} variable are randomly permuted in the OOB samples and accuracy computed again. The decrease in accuracy as a result of permuting is averaged over all trees and used as an importance measure of variable j (Hastie et al., 2008).

The mean decrease in gini measure is constructed from the decrease in node impurities from splitting on variable j. At each split in each tree, the improvement in the split-criterion is the importance measure that is attributed to the splitting variable. This is accumulated over all the trees in the forest separately for each variable (Hastie et al., 2008).

3.4 Receiver operating characteristic (ROC) curve

A Receiver operating characteristic (ROC) curve is the most commonly used way to visualize the performance of binary classifiers. It is a plot of the true positive rate (sensitivity) on the y-axis versus the false positive rate (1-specificity) on the x-axis for every possible classification threshold/cut-point (Park et al., 2004).

The True Positive Rate (sensitivity) aims to answers the question that, "When the actual classification is positive (adverse event detected), how often does the classifier predict positive?" while the False Positive Rate (1-specificity) answers the question that, "When the actual classification is negative (no adverse event detected), how often does the classifier incorrectly predict positive?".

In our study we aim at comparing the performance of different classifiers so area under the curve (AUC) which is a measure associated with the roc curve was used (Park et al., 2004). AUC measures the overall performance of a classifier and it can be thought to represent the probability that a classifier will rank a randomly chosen positive (AE present) observation higher than a randomly chosen negative (no AE) observation in terms of the probability.

Therefore, ROC curve and its attributes were used in the comparison of the performance of the different classification approaches employed in our study.

3.5 Software

`R version 3.2.1` was used for all statistical analyses and some of the packages used were; `randomForest, tree, stats, pROC`, and `grpreg`. All tests were done at the 5% level of significance.

4 Results

4.1 Comparison of characteristics

A total of 830 patients with an unplanned need for higher level of care had their records reviewed whereby in 465 (56%), one or more adverse events were detected and 365 (44%) had no adverse events. The the patients' age ranged from 21-95 years with majority 328 (39.5%) of the patients being in the 66-79 age category, followed by 239 (28.8%) 80+, 218 (26.3%) 41-65, and the smallest category was the 21-40 age group with 45 (5.4%). The length of hospital stay from admission to time of the unplanned transfer to higher level of care ranged from less than a day to 211 days with a median time of approximately 4 days.

A more detailed look in-to these characteristics is reported in the Appendix including; a cross-tabulation of the patient characteristics and their adverse event status is as reported on Tables (3 and 4) in the Appendix for the categorical characteristics. Table 5 in the Appendix shows the summary statistics of the polypharmacy, number of disciplines and the length of hospital stay from admission to time of transfer (LOS) whereby patients with an AE were observed to have a higher mean length of stay ($\approx$ 10 days) as compared to those without an AE ($\approx$ 8 days).

The process of model building was done using the training set and Table 6 in the Appendix reports the results for the univariate logistic regression analysis where the aim was to find out which variables independently show an association with the response at a pvalue of <0.25, as suggested by Bendel and Afifi (1977).

4.2 Logistic Regression Approach

4.2.1 Logistic regression - Stepwise approach

All clinical characteristics with a pvalue < 0.25 from the univariate analysis were considered as candidates of the important predictors and were considered in the multiple regression model. Table 1 shows the final multiple regression which was achieved through stepwise model selection approach. None of the Interactions were significant at α of 0.05 hence they were dropped from the final model. Although Age category and ASA (score of severity of illness) were not significant, they were retained in the model since they were of clinical importance.

Table 1: Results of Fitting the Multivariable Model

Effect	Odds Ratio	[95% CI]		
Intercept	0.735	0.230	2.354	
ASA (severity score): vs ASA 1				
ASA 2	0.745	0.287	1.934	
ASA 3	0.871	0.324	2.339	
ASA 4	0.792	0.298	2.101	
age category : vs 21-40				
41-65	2.117	0.793	5.656	
66-79	2.256	0.814	6.257	
80+	2.230	0.789	6.303	
previous hospitalization : vs No				
Yes	1.686	1.123	2.532	*
admission type: vs surgical trauma-unplanned				
medical - planned	0.516	0.217	1.228	
medical - unplanned	0.263	0.131	0.526	***
surgical - planned	2.529	1.036	6.176	*
Number disciplines	1.591	1.141	2.219	**

*Significance codes: **

From Table 1, it was observed that all age categories (older ages) were more likely to have an adverse event as compared to the younger patients (21-40 age category) while controlling for the other factors in the model but this difference was not significant. The number of disciplines the patient was treated by before transfer, the type of admission (either medical or surgical) and whether they had a previous admission within 3 months were all significant. For instance, those who had a previous admission were 1.69 times more likely to have an AE as compared to those without a previous admission, adjusting for the other covariates in the model.

The odds of getting an adverse event for patients with both planned and unplanned medical admission were, respectively, 0.52 and 0.26 times lower as compared to patients with an unplanned surgical trauma admission. Further, for every increase in the number of disciplines a patient was treated with the odds of an adverse event being present increased by a factor of 1.59.

From the patient characteristics collected from the time of admission to the time of the unplanned transfer to higher level of care, the following were selected as important predictors for the presence or absence of an adverse event;- the patients age category, score of severity of illness at admission to the hospital (ASA), type of admission (either medical or surgical), previous admission status, and the number of disciplines the patient was treated by before the transfer to a higher level of care.

Using these selected predictors, this model was refit on a 100 different sampled sets of train and test data. This was enable get a measure of the variability and performance of the model chosen. Table 2 reports the optimum threshold, specificity, sensitivity and the area under the under the roc which were used to assess the performance of the different classifiers.

4.2.2 Logistic regression - LASSO penalty

Logistic regression model was also fit containing all the possible p predictors, and the lasso shrinkage technique was used to shrink some of the regression coefficient to zero in a form of variable selection. This was done on the training set and Figure 4(a) in the Appendix shows the value of the tuning/shrinkage parameter, λ, which was selected through 10 fold cross-validation whereby the λ value and the set of predictors that gave the smallest misclassification/prediction error were selected.

The set of predictors whose coefficients were not shrunk to zero were selected as the most important predictors for the presence or absence of an adverse event and they included; score of severity of illness at admission (ASA), age category, previous hospitalization$<=$3 months, type of admission (either medical or surgical), number of disciples the patient was treated with before transfer, home living situation, whether the patient had an invasive treatment before transfer, surgery before the transfer, and the hour of admission.

To assess the performance of the model, 100 different sets of test and train data were randomly drawn whereby each time the model was fit using the train data and predictions done on the test set. Table 2 reports the measures used to assess the performance of the model in terms of the accuracy of the prediction and classification.

4.3 Tree Based

4.3.1 Classification Trees

The classification trees were grown using the `tree` function from `R` statistical programming language. Default criteria in the `tree` package were adopted where; at every node the split that maximize the reduction in deviance was chosen and smallest permitted node size was 10, thirdly, the node was not subsequently split if the within node deviance was $<$ 0.01 of that of the root node. Pruning was also done on the initial tree where the optimal number of leaves was determined through a 10 fold cross-validation to identify the the tree size with the smallest misclassification error as shown by Figure 4(b) in the Appendix.

After pruning, we grew a tree size of 5 terminal nodes and the variables used in the tree construction were; type of admission (medical vs surgical), previous hospitalization<=3 months, Length of stay, admission diagnosis and number of medications per day (polypharmacy).

From the tree (Figure 1), type of admission either medical or surgical (*med vs surg*) was selected as the most important factor in determining the presence of an AE. Patients with surgical (both planned and unplanned trauma) admissions were at a much higher risk of an adverse event (AE). Those with medical (both planned and unplanned) admission were further split according to their length of stay before transfer to higher level, those with >=10.57 days were more likely to have an AE. Patients with length of stay <10.57 days were further split depending on the polypharmacy whereby those with <16.5 had a higher probability of an AE and were stratified further according to the previous admission status. Those patients who did not have a previous admission within the past 3 months were split according to their admission diagnosis whereby those with either diseases of the digestive system, or of the skin and subcutaneous tissue or diseases of the musculoskeletal system and connective were found to be less likely to have an AE.

Figure 3(a) shows ROC curve and the area under the roc curve which was used in the assessing the accuracy of the predictions which are reported in detail in Table 2.

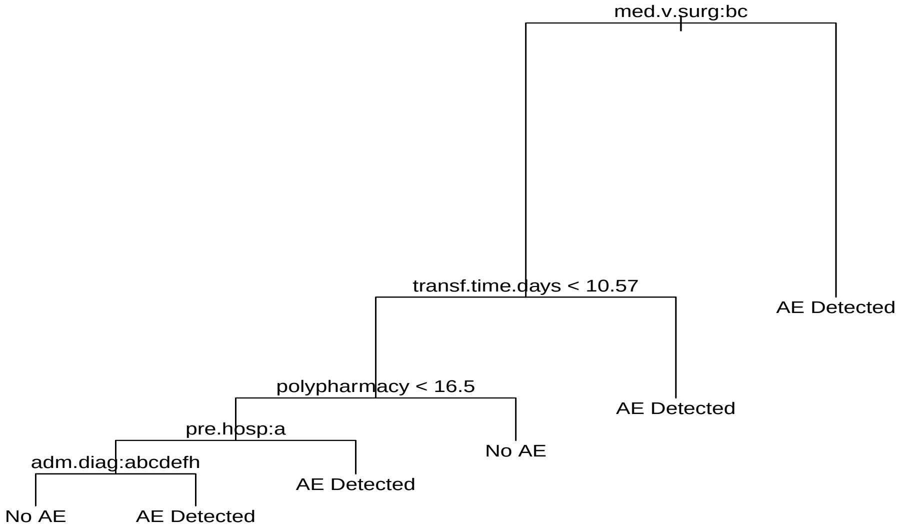


Figure 1: Classification Tree

4.3.2 Random Forest

Using the `randomForest` package in `R`, we grew random forests consisting of 1000 classification trees and set the number of variables chosen at each split (*mtry*) of each tree to be 3. This was the value that produced the lowest out of bag error as shown in Figure 5 in the Appendix. A training set (60% of the data) was used to train the random forest model. Predictions and classifications was done on the evaluation/test data set and were obtained by majority vote across the classification trees. The performance of the prediction is shown by the roc curve on Figure 3(b).

Figure 2 shows the two measures of the variable importance. The first measure was the decrease in accuracy which measures the prediction strength of each of each variable. The left plot on Figure 2 shows the ranked measure of the variable with type of admission (either medical or surgical) of admission showing the greatest prediction strength.

For the second measure, the importance of the variables was assessed by how much they contributed to reduction of the gini index (node impurity) whereby a higher decrease in gini means that a particular predictor variable played a greater role in partitioning the data into AE and no AE classes. The right plot on Figure 2 shows the ranked measure with length of stay from admission to transfer showing the greatest overall decrease in node impurity.

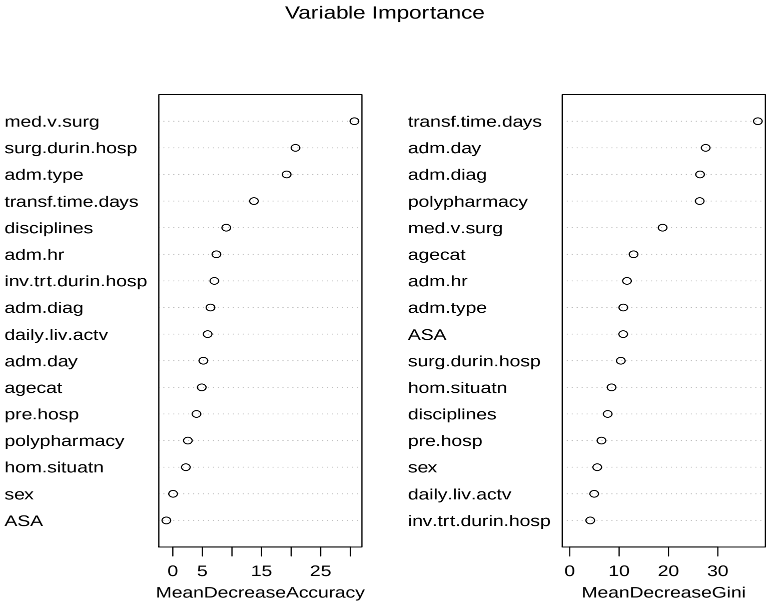
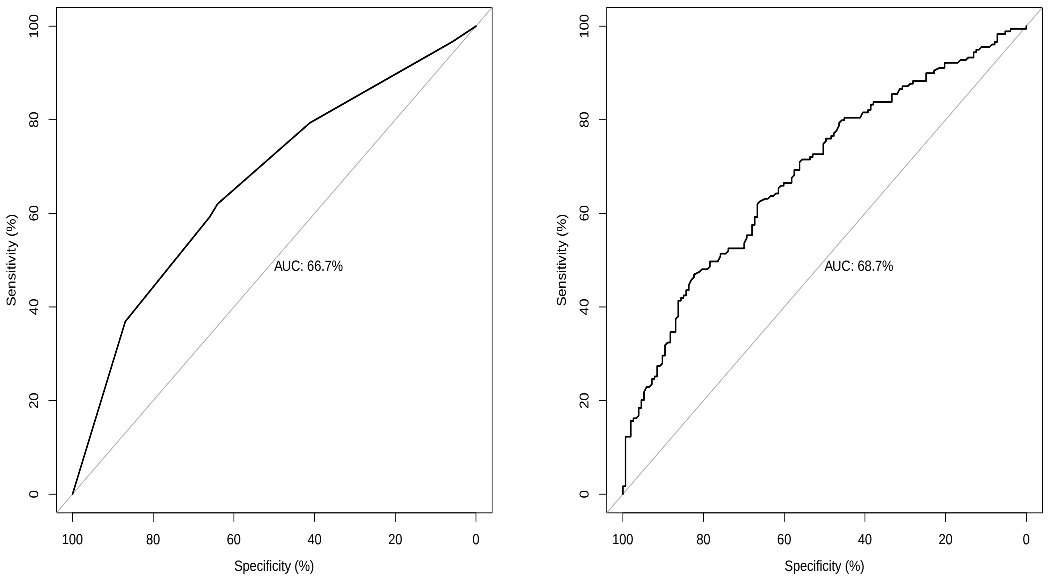


Figure 2: Variable importance plot for the random forest



(a)

(b)

Figure 3: ROC from the (a) Classification tree predictions (b) random Forest

4.4 Comparison of the Prediction and classification

The accuracy of the different classifiers depends on how well the different models can separate the patients into those with and without an adverse event. This accuracy of the prediction on the validation set was measured by comparing the area under the ROC curve (AUROC) for the different approaches. As reported in Table 2 the AUROC ranged from a low of 66.65% for the classification tree to 68.98% for the stepwise logistic regression which showed the greatest predictive accuracy for the probability of an adverse event. This implies that, the probability that our model will rank a randomly chosen patient who has an AE higher than a randomly chosen patient without an AE is 68.98%.

Table 2: Accuracy of the classification and prediction

	Logistic Approach		Tree Based	
	Stepwise approach[95% CI]	Lasso penalty[95% CI]	Classification tree	Random forest
Cutpoint	0.58[0.56, 0.59]	0.59[0.57, 0.60]	0.55	0.66
Specificity	77.93[76.45, 79.41]	77.08[75.39, 78.78]	64.05	82.35
Sensitivity	53.78[52.15, 55.41]	53.02[51.10, 54.93]	62.01	46.93
NPV	57.21[56.50, 57.91]	56.62[55.96, 57.28]	59.04	57.01
PPV	76.24[75.33, 77.15]	75.52[74.58, 76.46]	66.87	75.68
AUROC	**68.98**[68.54, 69.41]	**67.47**[67.06, 67.88]	**66.65**	**68.70**
Correct Classification	64.34[63.89, 64.80]	63.63[63.14, 64.12]	62.95	63.25

NPV-negative predictive value; PPV-positive predictive value
[95% CI] - Bootstrapped Confidence Intervals from the 100 samples

The accuracy of classification on the validation set for the different methods is also reported in Table 2 through the sensitivity, specificity, positive predictive value (PPV) and the negative predictive value (NPV). The sensitivity ranged from 46.93% for the random forest to a high of 62.01% for the classification tree. The specificity ranged from 64.05% for the classification tree to a high of 82.35% for the random forest. Logistic regression with stepwise selection approach produced a better specificity-sensitivity (77.93% - 53.78%) trade off and had the highest (64.34%) correct classification hence it was selected since it perform better compared to the other classifiers.

The random forest showed an improvement to the classification tree by increasing the proportion of correctly classified patient to 63.25% from 62.95%. There was also an increase in the classification accuracy for the patients without an adverse event in that the specificity increased to 82.35% but at an expense of the sensitivity which reduced to 46.93%.

From the stepwise logistic regression, the mean specificity was 77.93 implying that on average 77.93% of patients without an adverse event will be correctly identified as not having an adverse effect. In our case this can be viewed as the cases that do not need to be reviewed hence saving resources that would have otherwise been spent on them. The mean sensitivity was 53.78% which implies that 53.78% of the patients with an adverse effect will be correctly classified as having an adverse event.

5 Discussion

Unplanned transfers to higher level of care associated with adverse events has an impact on the patient in that it prolongs their length of hospital stay increasing the hospitalization cost which is a burden to both the patient and the family. These transfers also place additional pressure on the ICU resources and apart from the prolonged in-hospital stay the level of harm of the transfers can range from temporary harm, longterm/permanent impairment, and even death (Marquet et al., 2015).

Retrospective medical record review is currently one of the commonly used and the best methods available to assess the prevalence of adverse events (Garry et al., 2014). Despite its positive impact, a shortcoming of this approach is the fact that it is costly and time consuming.

Resource limitations often restrict the ability of public health personnel in health interventions aimed at identifying and eliminating health issues hence modeling approaches can help reducing this burden. In identification; statistical modeling will help identify important patient characteristics that the clinicians should keep a keen eye on hence increases the likelihood of reducing the occurrence of the AEs. In elimination; coming up with models than can correctly classify a patient can reduce the workload in the medical review process by identifying records to dedicate resources to.

In this study we report several models for predicting and classifying adverse events in patients with an unplanned transfer to higher level of care. Specifically, we used supervised learning methods, which include regression based methods (e.g., generalized linear models, and penalized regression) and tree based methods (including, decision trees and random forests). The performance of these approaches in terms of the accuracy in prediction and classification was evaluated through the area under the roc curve, sensitivity and specificity.

A study by Austin et al. found out flexible tree-based methods, random forests, offered substantial improvement in prediction and classification to the classification trees but convectional logistic regression was able to more accurately predict the probability of the event of interest which is also reflected in this analysis.

Even though the classification tree give a simple intuitive interpretation since they closely mirror decision making steps, they fall short in terms of their predictive accuracy (James et al., 2013) as seen in our analysis. Random forest showed an improvement to the classification approach by increasing the accuracy of prediction to 68.70% (from 66.65%) but this is at an expense of the ease of interpretation of the resulting model. There was also an improvement in the specificity to 82.35% but at an expense of lower specificity of 46.93% hence the specificity-sensitivity trade off was not as good as compared to the logistic approaches.

Through the logistic regression approaches, the stepwise logistic model attained a mean specificity of 77.93% implying that out of the patients without an adverse event 77.93% will be correctly identified. This points to the proportion of patients records that the health workers do not require to review. This approaches also had the highest prediction accuracy (auroc of 68.98%). The model had a PPV of 76.24% meaning that if the model predicts an AE, in 76.24% of the cases this is true. So if the model is used to pre-select a number of records, its expected to find an AE in 76.24% of the records which is an improvement on the efficiency of the search in that the randomly selected records from the record review only had a prevalence of 56%.

From the stepwise logistic model it was observed that, older patients were more likely to have an adverse event as compared to the younger patients while controlling for the other factors in the model. Patients who had a previous admission were also more likely to have an AE as compared to those without a previous admission. Patients with both planned and unplanned medical admission had a lower probability of an adverse event as compared to patients with an an unplanned surgical trauma admission. Further, for every increase in the number of disciplines a patient was treated with the odds of an adverse event being present increased.

An advantage of the logistic approaches is that they generate probabilities of class membership for each patient.Whereas the tree based approaches only generates average probabilities applicable to all the patients assigned to a particular group. Also, since the tree based approach are intended to identify distinct population subgroups, their hierarchical nature does not allow the estimation of the net impact of a single independent variable (Stephan and Lucila, 2002). So its useful in situations where objective is not to test the hypothesis of the independent variables.

In summary, some of the patient characteristic that may be used as reliable indicators/predictors of an adverse event as suggested by the stepwise logistic model include; type of admission (medical or surgical), whether the patient had a previous hospitalization $\leq$3 months, score of severity of illness at admission (ASA), age category, and number of disciplines the patient was treated by before the transfer to a higher level. In using these predictors, the work load on review process may be reduced through classification by guiding on the proportion of records the health workers do not need to review.

Bibliography

Austin, P.C., Tu, J.V., Ho, J.E., Levy, D., and Lee, D.S. (2013), "Using methods from the data mining and machine learning literature for disease classification and prediction: A case study examining classification of heart failure sub-types." *Journal of clinical epidemiology*, 66(4), 398–407.

Baker, G.R., Norton, P.G., Flintoft, V., Blais, R., Brown, A., J. Cox, ..., and Tamblyn, R. (2004), "The canadian adverse events study: the incidence of adverse events among hospital patients in canada." *Canadian Medical Association Journal*, 170(11), 1678–1686.

Bendel, R.B. and Afifi, A.A. (1977), "Comparison of stopping rules in forward "stepwise" regression." *Journal of the American Statistical Association*, 72(357), 46–53.

Breiman, Leo, Friedman, Jerome, Stone, Charles J, and Olshen, Richard A (1984), *Classification and regression trees*. CRC press.

Dasgupta, A., Sun, Y.V., König, I.R., Bailey-Wilson, J.E., and Malley, J.D. (2011), "Brief review of regression-based and machine learning methods in genetic epidemiology: The genetic analysis workshop 17 experience." *Genetic Epidemiology*, 35(Suppl 1), S5–11.

de Vries, EN, Ramrattan, MA, Smorenburg, SM, Gouma, DJ, and Boermeester, MA (2008), "The incidence and nature of in-hospital adverse events: A systematic review." *Qual Saf Health Care*, 17(3), 216–223.

Garry, D.A., McKechnie, S.R., Culliford, D.J., Ezra, M., Garry, P.S., Loveland, RC, Sharma, V.V., Walden, A. P., and Keating, L.M. (2014), "A prospective multicentre observational study of adverse iatrogenic events and substandard care preceding intensive care unit admission (prevent)." *Anaesthesia*, 69(2), 137–142.

Hastie, T., Tibshirani, R., and Wainwright, M (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome (2008), *The elements of statistical learning: data mining, inference and prediction*, 2 edition. Springer, URL `http://scholar.google.com/scholar.bib?q=info:roqIsr0iT4UJ:scholar.google.com/&output=citation&hl=en&ct=citation&cd=0`.

Hosmer, DAVID W. and Lemeshow, STANLEY (2000), *Applied Logistic Regression (2nd edition)*. John Wiley & Sons, New York.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics, Springer New York, URL `https://books.google.com/books?id=at1bmAEACAAJ`.

Kurt, Imran, Ture, Mevlut, and Kurum, A. Turhan (2008), "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease." *An International Journal*, 34, 366–374.

Marquet, K., Claes, N., Troy, E. De, Kox, G., Droogmans, M., ..., W. Schrooten, and Vleugels, A. (2015), “One fourth of unplanned transfers to a higher level of care are associated with a highly preventable adverse event: A patient record review in six belgian hospitals.” *Critical Care Medicine*, 43(5), 1053–1061.

Mercier, E., Giraudeau, B., Giniès, G., Perrotin, D., and Dequin, P.F. (2010), “Iatrogenic events contributing to icu admission.” *Intensive Care Med*, 36(6), 1033–1037.

Michel, Philippe (2003), “Strengths and weaknesses of available methods for assessing the nature and scale of harm caused by the health system: literature review. geneva: World health organization (who).” URL `http://www.who.int/patientsafety/research/P_Michel_Report_Final_version.pdf`.

Park, S. H., Goo, J. M., and Jo, C.-H. (2004), “Receiver operating characteristic (roc) curve: Practical review for radiologists.” *Korean Journal of Radiology*, 5(1), 11–18.

Posa, P.J., Yonkee, D.E, and Fields, W.L (1992), “Development and implications of an interdisciplinary quality assurance monitor on unplanned transfers into the intensive care units.” *J Nurs Care Qual*, 6, 51–55.

Soop, M., Fryksmark, M., U.and Köster, and Haglund, B. (2009), “The incidence of adverse events in swedish hospitals: a retrospective medical record review study.” *International Journal for Quality in Health Care*, 21(4), 285–291.

Stephan, Dreiseitl and Lucila, Ohno-Machado (2002), “Logistic regression and artificial neural network classification models: a methodology review.” *Journal of Biomedical Informatics*, 35(5-6), 352–359.

Van den Heede, K, Sermeus, W, Diya, L, Lesaffre, E, and Vleugels, A (2006), “Adverse outcomes in belgian acute hospitals: retrospective analysis of the national hospital discharge dataset.” *International Journal for Quality in Health Care*, 18(3), 211–219.

Wilson, R.M., Runciman, W.B., and R.W. Gibberd, et al (1995), “The quality in australian health care study.” *Med J Aust*, 163(9), 458–471.

Worster, A. and Haines, T. (2004), “Advanced statistics: Understanding medical record review (mrr) studies.” *Academic Emergency Medicine*, 11, 187–192.

6 Appendix

Table 3: Crosstabulation of patient characteristics and the adverse event status

	Adverse Events	
Patient characteristic	**No AE** (%)	**AE Detected** (%)
sex		
M	180(0.43)	241(0.57)
F	185(0.45)	224(0.55)
age Category		
21-40	27(0.60)	18(0.40)
41-65	89(0.41)	129(0.59)
66-79	142(0.43)	186(0.57)
80+	107(0.45)	132(0.55)
ASA class (score of severity of illness)		
ASA 1	25(0.42)	35(0.58)
ASA 2	76(0.44)	95(0.56)
ASA 3	88(0.38)	143(0.62)
ASA 4	176(0.48)	192(0.52)
previous hospitalization <=3months		
no	215(0.48)	235(0.52)
yes	150(0.39)	230(0.61)
type of admission		
elective	61(0.26)	172(0.74)
emergency	278(0.52)	260(0.48)
transfer from other hospital	6(0.26)	17(0.74)
admission after consultation	20(0.56)	16(0.44)
medical vs surgical type		
medical - planned	33(0.41)	48(0.59)
medical -unplanned	286(0.56)	227(0.44)
surgical - planned	24(0.15)	131(0.85)
surgical trauma - unplanned	22(0.27)	59(0.73)
home situation/group living situation		
home living alone	88(0.50)	89(0.50)
home cohabiting	242(0.41)	342(0.59)
staying in institution	30(0.55)	25(0.45)
not known	5(0.36)	9(0.64)
correct ward		
yes	339(0.44)	438(0.56)
no	26(0.49)	27(0.51)
activities of daily live		
no limitations	121(0.45)	164(0.55)
disfunction	244(0.45)	301(0.55)
cognitive impairment		
no	327(0.43)	426(0.57)
yes	38(0.49)	39(0.51)

Table 4: Table 3 Continued . . .

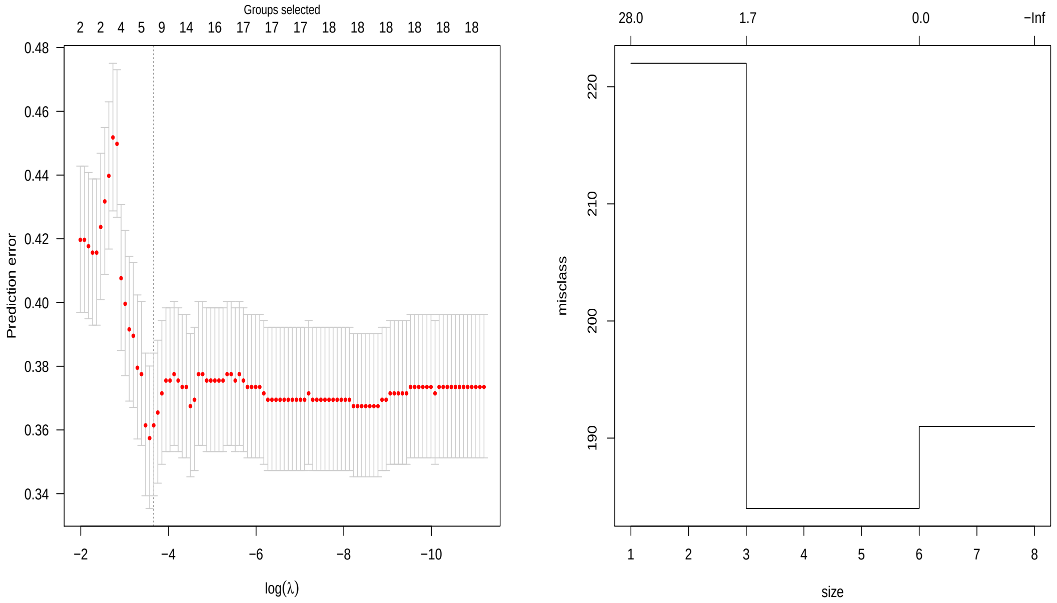
	Adverse Events	
Patient characteristic	**No AE**	**AE Detected**
surgery before the transfer		
no	289(0.54)	247(0.46)
yes	76(0.26)	218(0.74)
invasive treatment before transfer		
no	321(0.45)	397(0.55)
yes	44(0.39)	68(0.61)
DNR (do not reanimate) code		
no	350(0.44)	453(0.56)
DNR1	7(0.50)	7(0.50)
DNR2	8(0.62)	5(0.38)
admission in the hospital: week vs weekend		
week	292(0.43)	387(0.57)
weekend	73(0.48)	78(0.52)
day of admission		
Monday	65(0.40)	97(0.60)
Tuesday	56(0.41)	79(0.59)
Wednesday	60(0.49)	63(0.51)
Thursday	57(0.41)	81(0.59)
Friday	54(0.45)	67(0.55)
Saturday	34(0.52)	32(0.48)
Sunday	39(0.46)	46(0.54)
hour of admission		
evening(16-23:59)	117(0.51)	113(0.49)
day(8-15:59)	191(0.44)	240(0.56)
overnight(00-07:59)	57(0.34)	112(0.66)
admission diagnosis: List of ICD-9 codes		
011-139, 240-279: Infectious, Endocrine	14(0.40)	21(0.60)
140-239, 280-289: Neoplasm, disease of blood	25(0.31)	56(0.69)
290-319, 320-359, 360-389: Mental, Nervous system, sense organs	18(0.47)	20(0.53)
390-459: diseases of the circulatory system	116(0.51)	110(0.49)
460-519: disease of the respiratory system	72(0.60)	48(0.40)
520-579: diseases of the digestive system	40(0.38)	66(0.62)
580-629, 630-679: genitourinary system, pregnancy complication	20(0.49)	21(0.51)
680-709, 710-739: disease of Skin, musculoskeletal	8(0.22)	28(0.78)
780-799, 800-999, E & V, not classified	52(0.35)	95(0.65)

Table 5: Table 3 Continued . . .

	Adverse Events	
Patient characteristic	**No AE**	**AE Detected**
polypharmacy (number of medications/day)		
mean	7.72	7.12
median	7.00	7.00
sd	4.98	4.35
range	[0.00, 23.00]	[0.00, 20.00]
Number of disciplines		
mean	1.28	1.40
median	1.00	1.00
sd	0.65	0.68
range	[1.00, 5.00]	[1.00, 5.00]
LOS in ward before transfer		(days)
mean	8.42	10.39
median	2.86	4.35
sd	18.04	18.43
range	[0.05, 180.5]	[0.05, 211]

Table 6: Results of Fitting Univariable Logistic Regression

Variable	Description	Deviance	pvalue
sex	sex	1.962	0.161
agecat	age category	2.040	0.564
ASA	score of severity of illness	1.115	0.774
pre.hosp	previous hospitalization <=3months	6.971	0.008
polypharmacy	number of medications/day	1.966	0.161
hom.situatn	home situation/ grp living situation	6.616	0.085
adm.type	type of admission	30.901	<.0001
med.v.surg	medical vs Surgical	68.415	<.0001
correct.ward	correct ward	2.046	0.153
disciplines	number of disciplines	9.638	0.002
daily.liv.actv	activities of daily live	0.000	0.988
cog.impair	cognitive impairment	1.133	0.287
surg.durin.hosp	surgery before the transfer	42.322	<.0001
inv.trt.durin.hosp	invasive treatment before transfer	0.114	0.736
DNR	do not reanimate code	2.082	0.353
transf.time.days	LOS in ward before transfer	5.645	0.018
weeknd.week	admission in the hospital: week vs weekend	2.139	0.144
adm.day	day of admission in the hospital	6.036	0.419
adm.hr	hour of admission in the hospital	9.542	0.008
adm.diag	diagnosis at admission	23.240	0.003



(a) (b)

Figure 4: Cross-validation for choosing the tuning parameter (a) choice of λ for the group lasso penalty (b) pruning parameter for the classification tree

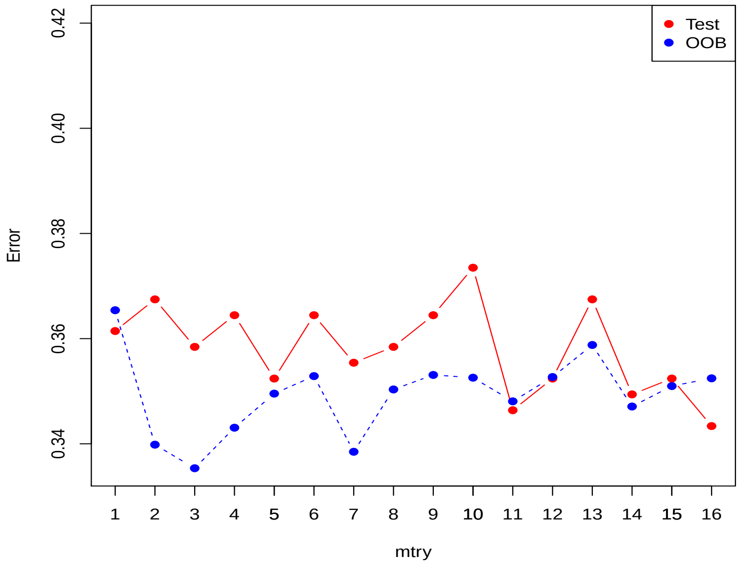


Figure 5: Choosing the tuning $(mtry)$ parameter for the random forest.

R-Codes

```
setwd("F:/4.Thesis/Codes")

# Read in the data
review<- readRDS("rev.rds") #
attach(review)
#===========================================================================#
#  PREDICTION & CLASSIFICATION
# 1) LOGISTIC REGRESSION APPROACHES
#===========================================================================#
#_________________________________________________________________________
# 1.1) Ordinary Logistic Regression
# Univariate Logistic reg
#_________________________________________________________________________
# sampling and Spliting the data into train:test - 60:40
set.seed(37)
train<-sample(1:nrow(review),(0.6*nrow(review)))
test<- -train
train.data<-review[train,]
test.data<-review[test,]
# list of the potential predictors
varlist<-c("sex","agecat","ASA","pre.hosp","polypharmacy","hom.situatn",
           "adm.type","med.v.surg","correct.ward","disciplines",
           "daily.liv.actv","cog.impair","surg.durin.hosp","inv.trt.durin.hosp"
           ,"DNR","transf.time.days","weeknd.week","adm.day","adm.hr","adm.diag")

# Fitting Univariate Logistic Regression
# looping Over all the univariate models and Fiting a Logistic reg
library(aod)
library(stats)
models <- lapply(varlist, function(x) {
    glm(substitute(AE ~ i, list(i = as.name(x))), data = train.data
    , family = "binomial")
})
# Functions to extract variable name
my.var<-function(x){
    row.names(anova(x,test="Chisq"))[2]
}
# Functions to extract the deviance
my.Dev<-function(x){
    (anova(x,test="Chisq"))$Deviance[2]}
# Functions to extract the overal pvalue
my.pval<-function(x){
    (anova(x,test="Chisq"))$`Pr(>Chi)`[2]
}
# Objects to collect some information
variable<-NULL; Deviance<-NULL;pvalue<-NULL
```

```
# looping Over all the univariate models
for (i in 1 :length(models)){
    variable[i]<- lapply(models,my.var)[i]
    Deviance[i]<- lapply(models,my.Dev)[i]
    pvalue[i]<- lapply(models,my.pval)[i]
}
# All the patient characteristics and their p-values from the univariate analysis.
(uni.model<-cbind(variable,Deviance=round(as.numeric(Deviance),4)
,pvalue=round(as.numeric(pvalue),5)))
# cutpoint of ‘pvalue<=0.25‘ the following list displays
# variables that shows at least some association with the outcome were as follows.
(varlist.uni<-as.character(uni.model[,1][uni.model[,3]<=0.25]))
#___________________________________________________________________
# Fitting Multiple Logistic reg
#___________________________________________________________________
summary(M1<- glm(AE~ agecat + ASA + sex + pre.hosp + polypharmacy + hom.situatn +
adm.type + med.v.surg + correct.ward + disciplines +  weeknd.week +
                     surg.durin.hosp + transf.time.days + adm.hr + adm.diag
                 ,data = train.data, family = "binomial")) #
# Test statistic for model fit
# Ho : all coefficients = 0
# H1 :  AT least one of the coefficient !=0
# P-value shows  current fits better than null
with(M1, pchisq(null.deviance-deviance, df.null-df.residual,lower.tail= FALSE))
#____
# Variable selection Methods. Stepwise model selection techniques
# Using the ‘step‘ function that is based on the AIC and
# the procedure stops when the AIC criterion cannot be improved.
search <- step(M1,~.)
## summary of the search
search$anova
# The final main effects model chosen is as follows.
# ASA and Agecat were not significant but were maintained in the model
summary(M2<-glm(AE ~ ASA + agecat + pre.hosp + med.v.surg + disciplines
                ,data = train.data, family = "binomial"))
# Test statistic for model fit
with(M2, pchisq(null.deviance-deviance, df.null-df.residual,lower.tail= FALSE))
## Checking for all possible interactions
# Some combinations lead to 0 cell counts
search <- step(M2, ~.^2)
# summary of the search
search$anova
# none of interactions were significant
summary(Mod.2.1<-glm(AE ~ pre.hosp + med.v.surg + disciplines +
med.v.surg:disciplines,data = train.data, family = "binomial"))
#____________________________________________
# Final Model..
#____________________________________________
```

```
#
summary(M2<-glm(AE ~ ASA + agecat + pre.hosp +
med.v.surg + disciplines,data = train.data, family = "binomial"))
# OR and their 95% CI
exp(cbind(OR = coef(M2), confint.default(M2)))
# Diagnostics and the plots
glm.diag(M2)
glm.diag.plots(M2)
#_____________________
#Variability of the predictions,pecificity,Sensitivity,AUC and optimal cutpoint.
#______
# To get the variability in the prediction performance of the models
# on the test data,
# the models were refit 100 times on different sampled train data
# and prediction on the test data assesed.
library(pROC)
library(caret)

threshold<-NULL;specificity<-NULL;sensitivity<-NULL;NPV<-NULL ;PPV<-NULL
cls.tst<-NULL;auc.logit<-NULL
# Drawing 100 different samples
set.seed(37)
for (i in 1:100){
    train<-sample(1:nrow(review),(0.6*nrow(review)))
    test<- -train
    train.data<-review[train,]
    test.data<-review[test,]
# AE grouping for testing datatset
    tst.AE<-AE[test]
# model fit on train
    M2.2<-glm(AE ~ ASA + agecat + pre.hosp + med.v.surg + disciplines
              ,data = train.data, family = "binomial")
# predicted probs on test data
    prd.M2.2.test<-predict(M2.2,test.data,type="response")
# roc curve attributes
    roc<-data.frame(tst.AE,prd.M2.2.test)
    names(roc)<-c("t","pd")
    roc(roc$t,roc$pd)
    roc(t~pd,roc)
    roc1 <- roc(roc[,1],roc[,2], percent=TRUE)
    r<-coords(roc1, "best", ret=c("threshold", "specificity",
    "sensitivity","npv","ppv"))
    threshold[i]<-r[1];specificity[i]<-r[2];
    sensitivity[i]<-r[3];NPV[i]<-r[4] ;PPV[i]<-r[5]
# Area Under the ROC
    auc.logit[i]<-auc(roc1)
# Changing the predicted probabilities into to classes
# using optimal cutpoint(threshold) from the roc function
# classifications On the test data
    test.cl<-rep("No AE", dim(test.data)[1])
    test.cl[prd.M2.2.test>threshold[i]]<-"AE Detected"
# correct classification rate.
    cls.tst[i]<-mean(test.cl == tst.AE)
}
library(Rmisc)
multi.fun=function(x){
    c(mean=mean(x),median=median(x),range=range(x),CI=CI(x, ci = 0.95)
    )}
# Optimal Threshold.
round(multi.fun(threshold),3)
# Specificity.
round(multi.fun(specificity),3)
# Sensitivity.
round(multi.fun(sensitivity),3)
# NPV
round(multi.fun(NPV),3)
# ppv
round(multi.fun(PPV),3)
# Area under the ROC
round(multi.fun(auc.logit),3)
# Predicted Correct classification on the Test Data
round(multi.fun(cls.tst)*100,3)
#===============================================================================#
#  1.1) Group LASSO Penalty Approach
#_____________________________________________________________________________
library(grplasso)
library(grpreg)
varlist<-c("AE","sex","agecat","ASA","pre.hosp","polypharmacy","hom.situatn",
           "adm.type","med.v.surg","correct.ward","disciplines", "adm.diag",
           "daily.liv.actv","cog.impair","surg.durin.hosp","inv.trt.durin.hosp",
           "DNR","transf.time.days","weeknd.week","adm.day","adm.hr")
r1<-review[,varlist]
x1=model.matrix(AE~.,r1)[,-1]
y1=r1$AE

set.seed(37)
train=sample(1:nrow(x1), (nrow(x1)*0.6))
test=(- train )
y.test=y1[test]

#Defining the groups.i.e defining Dummies from same variable to represent a group
group<-NULL
r2<-review[,varlist[-1]]
for(i in 1:ncol(r2)){
    if ( nlevels(r2[,i])>1){
        tmp<-  rep(i,(nlevels(r2[,i]))-1)
        assign(paste("n",i, sep = ""),tmp)
    } else {
        tmp<-i
        assign(paste("n",i, sep = ""),tmp )
    }
    group=c(group,tmp)
    i=i+1
}
#coefficients included in the model without being penalized, assigned to group 0
# agecat and ASA are maintained in the model
group[2:7]<-0
#_____________________
# Lambda value that minimizes the prediction error
# Plot showing a possible value of lambda, the number of Important variables
y.1<-as.numeric(y1)-1
fit.cv<-cv.grpreg(x1[train ,],y.1[train],group=group, nfolds=10,
trace=FALSE,family="binomial")
plot(fit.cv,type="pred") # plots the prediction error
# Variables with Non zero coefficients. Those Considered Important
cof.gp<-predict(fit.cv,x1[test,],lambda=fit.cv$lambda.min,type="coefficients")
[1:dim(x1)[2],]
cof.gp[cof.gp!=0]
#_____________________________________________________________________
#Variability of the predictions, specificity, sensitivity and optimum cutpoint.
#_________________
# Sampling and fitting the model 100 times to get the variablity
# of the predicted classification.
# The attributes of the ROC curve were also calculated
# and averaged out over the different samples drawn to also get their variation.

allvar<-NULL; correct.class.test<-NULL
threshold.glas<-NULL;specificity.glas<-NULL;sensitivity.glas<-NULL
NPV.glas<-NULL ;PPV.glas<-NULL
auc.glasso<-NULL

set.seed(37)
# looping 100 times
for (i in 1:100){
    train=sample(1:nrow(x1), (nrow(x1)*0.6))
    test=(- train )
    y.1<-as.numeric(y1)-1

fit <- grpreg(x1[train ,],y.1[train],group,penalty="grLasso", family="binomial")
# choosing the Tuning prameter(lambda) using 10 fold cross validation
    fit.cv<-cv.grpreg(x1[train ,],y.1[train],group=group, nfolds=10,
    trace=FALSE,family="binomial")
# predicted class on test data using 0.5 cutpoint
    class.test<-predict(fit.cv,x1[test,],lambda=fit.cv$lambda.min,type="class")
    # correct clasification
    correct.class.test[i]<-mean(class.test==y.1[test])
# Coefficients of the predictors
cof.gp<-predict(fit.cv,x1[test,],lambda=fit.cv$lambda.min,type="coefficients")
    [1:dim(x1)[2],]
    assign(paste("n",i, sep = "_"),cof.gp)
    allvar<-cbind(allvar,cof.gp)
    colnames(allvar)[i]<-(paste("iter",i, sep = "_"))
#  Predicted probability on test data
gp.pred=predict(fit.cv ,x1[test,],lambda=fit.cv$lambda.min,type="response")#
    roc.gl<-data.frame(t=y.1[test], pd =(gp.pred))
    roc.glas<- roc(roc.gl[,1],roc.gl[,2], percent=TRUE)
# roc curve attributes
    glas<-coords(roc.glas, "best", ret=c("threshold", "specificity",
    "sensitivity","npv","ppv"))
    threshold.glas[i]<-glas[1];specificity.glas[i]<-glas[2]
    ;sensitivity.glas[i]<-glas[3]
    NPV.glas[i]<-glas[4];PPV.glas[i]<-glas[5]
# Area under the curve
    auc.glasso[i]<-auc(roc.glas)
# Correct classification using optimal cutpoint
    t.cl<-rep("No AE", dim(x1[test,])[1])
    t.cl[gp.pred>threshold.glas[i]]<-"AE Detected"
#  Classification rate.using optimal cutpoint
    correct.class.test[i]<-mean(t.cl == AE[test])
}
#____
# A Distribution of the Classification probabilities
multi.fun=function(x){
c(mean=mean(x),median=median(x),range=range(x), CI=CI(x, ci = 0.95)
    )}
# Threshold
round(multi.fun(threshold.glas),3)
# Specificity
round(multi.fun(specificity.glas),3)
# Sensitivity
round(multi.fun(sensitivity.glas),3)
# NPV
round(multi.fun(NPV.glas),3)
# PPV
round(multi.fun(PPV.glas),3)
# Area under the ROC
round(multi.fun(auc.glasso),3)
# Predicted correct classification using the optimal cutpoint
round(multi.fun(correct.class.test)*100,3)
#____________________________________________________________________________________
# 2) TREE BASED Approaches
#____________________________________________________________________________________
```

```
# 2.1) CLASSIFICATION TREES
#____________________________
# splitting data into testing and training day
# 60% for training without replacement
set.seed(37)
train<-sample(1:nrow(review),(0.6*nrow(review)))
test<- -train
train.data<-review[train,] #498
test.data<-review[test,] #332
AE.test<-AE[test]

library(tree)
library(rpart)

varlist<-c("AE","sex","agecat","ASA","pre.hosp","polypharmacy","hom.situatn",
           "adm.type","med.v.surg","correct.ward","disciplines", "adm.diag",
           "daily.liv.actv","cog.impair","surg.durin.hosp","inv.trt.durin.hosp",
           "DNR","transf.time.days","weeknd.week","adm.day","adm.hr")
# fit using train data
t1<-tree(AE ~ ASA + sex + agecat + pre.hosp + polypharmacy + med.v.surg
         + surg.durin.hosp + transf.time.days+ hom.situatn + daily.liv.actv
         + cog.impair + adm.type + correct.ward + inv.trt.durin.hosp
         + DNR +  weeknd.week + adm.day + adm.hr + adm.diag + disciplines
         ,data = train.data)
summary(t1)
# Detailed version of the process
t1
# plot of the tree
plot(t1)
text(t1)
#__________________________________________
# Pruning
#__________________________________________
# bushy tree might have too  much variance
# using misclassification as the basis for prunning
# 10 fold cross-validation
set.seed(37)
cv.tr<-cv.tree(t1, FUN=prune.misclass)
# plot of misclass vs Size
# smallest miss class
plot(cv.tr)

# pruning the tree to size of 5
prune.tr<-prune.misclass(t1,best=5)
plot(prune.tr)
text(prune.tr)#pretty = 0

#______________________
```

```
# ROC curve for the Classification Tree
# using predicted probability (through class votes)of Adverse Event
t.pred<-predict(prune.tr,test.data, type="vector")[,2]

roc.CT<-data.frame(test.data$AE,t.pred)
names(roc.CT)<-c("t","pd")
roc(roc.CT$t,roc.CT$pd)
roc(t~pd,roc.CT)
roc.CT1 <- roc(roc.CT[,1],roc.CT[,2], plot=T,print.auc=TRUE,percent=T)

# Threshold, sensitivity, specificity, npv, ppv
(C.Tree<-coords(roc.CT1, "best", ret=c("threshold", "specificity",
"sensitivity","npv","ppv")))
# area under ROC
auc(roc.CT1)
#_____________________________________________________________________
#  2.2)     Random Forest
#_____________________________________________________________________
# More powerfull prediction for improving the prediction Accuracy of Trees
rf.list<-c("sex","agecat","ASA","pre.hosp","polypharmacy","hom.situatn",
           "adm.type","med.v.surg","correct.ward","disciplines", "adm.diag",
           "daily.liv.actv","cog.impair","surg.durin.hosp","inv.trt.durin.hosp",
           "DNR","transf.time.days","weeknd.week","adm.day","adm.hr")
set.seed(37)
train<-sample(1:nrow(review),(0.6*nrow(review)))
test<- -train
train.data<-review[train,] #498
test.data<-review[test,] #332
AE.test<-AE[test]

require(randomForest)
set.seed(37)
rf<-randomForest(as.factor(AE) ~  ASA + sex + agecat + pre.hosp + polypharmacy +
med.v.surg + disciplines + surg.durin.hosp + transf.time.days+ hom.situatn
+ daily.liv.actv + cog.impair + adm.type + correct.ward + inv.trt.durin.hosp
+ DNR +  weeknd.week + adm.day + adm.hr + adm.diag, ntree=1000,
data = train.data,importance=T)
rf
varImpPlot(rf,main="variable importance")
#------
#dropped some of the variables with less importance score
#up to the point the out of bag error drops compared to wen all var' included

set.seed(37)
rf.1<-randomForest(as.factor(AE) ~ ASA +  agecat + sex + pre.hosp + polypharmacy
+ med.v.surg + inv.trt.durin.hosp + disciplines + surg.durin.hosp
+ hom.situatn + adm.type + adm.day + adm.hr + adm.diag + daily.liv.actv
+ transf.time.days,ntree=1000, data = train.data,importance=T)
```

```
rf.1
varImpPlot(rf.1,main="variable importance")
#_________________________
# tuning parameter(mtry)- # of variables chosen at each split of each tree
# fitting a series of random forests. choose mtry when oob error is smallest
# mtry will range from 1:last variable

oob.err<-NULL;test.err<-NULL;corect.cl<-NULL
(L<-dim(rf.1$importance)[1])
set.seed(37)
for (mtry in 1:L){
fit=randomForest(as.factor(AE) ~  ASA +  agecat + sex + pre.hosp + polypharmacy
    + med.v.surg + disciplines + surg.durin.hosp + transf.time.days +
    hom.situatn + adm.type + adm.day + adm.hr + adm.diag + daily.liv.actv
    + inv.trt.durin.hosp,data = train.data,mtry=mtry,ntree=1000)
    oob.err[mtry]=mean((fit$err.rate)[,1])
    pred=predict(fit,test.data)
    test.err[mtry]=with(test.data,mean(as.character(pred)!=as.character(AE)))
    corect.cl[mtry]=with(test.data,mean(as.character(pred)==as.character(AE)))
    cat(mtry," ")
}

# Plot
matplot(1:mtry,cbind(test.err,oob.err),pch=19, col=c("red","blue"),type="b"
        ,ylab="Error",ylim=c(min(oob.err),0.42),xlab="mtry", xlim=c(1,L))
axis(1,las=1,at=1:L)
legend("topright",legend=c("Test","OOB"),pch=19, col=c("red","blue"))

#________________________________________
# Refiting using the tuning parameter
set.seed(37)
rf2<-randomForest(as.factor(AE)~ ASA+agecat+sex+pre.hosp+polypharmacy+med.v.surg
                  + disciplines + surg.durin.hosp + transf.time.days + hom.situatn
                  + adm.type + adm.day + adm.hr + adm.diag + daily.liv.actv +
                  inv.trt.durin.hosp
                  ,ntree=1000,mtry=3,data = train.data,importance=T) #
rf2
# A Dot plot of the variable importance measured by the Random Forest
varImpPlot(rf2,main="Variable Importance")
importance(rf2)
#_________________________
# predicted probability of Avderse Event
rf.pred<-predict(rf2,test.data, type="prob")[,2]
# ROC curve
roc.rf<-data.frame(test.data$AE,rf.pred)
names(roc.rf)<-c("t","pd")
roc(roc.rf$t,roc.rf$pd)
roc(t~pd,roc.rf)
```

```
roc.rf1 <- roc(roc.rf[,1],roc.rf[,2], plot=T,print.auc=TRUE,percent=T)
# Threshold, sensitivity, specificity,npv,ppv
(R.f<-coords(roc.rf1, "best", ret=c("threshold", "specificity",
"sensitivity","npv","ppv")))
# area under the curve
auc(roc.rf1)
#_________________________________________________________________
```