Evaluation of Random Forest and Ensemble Methods at Predicting Complications Following Cardiac Surgery

Abstract. Cardiac patients undergoing surgery face increased risk of postoperative complications, due to a combination of factors, including: higher risk surgery, their age at time of surgery and the presence of co-morbid conditions. They will therefore require high levels of care and clinical resources throughout their perioperative journey (i.e. before, during and after surgery). Although surgical mortality rates in the UK have remained low, postoperative complications on the other hand are common and can have a significant impact on patients' quality of life, increase hospital length of stay and healthcare costs. In this study we used and compared several machine learning methods – random forest, adaboost, gradient boosting model and stacking – to predict severe postoperative complications after cardiac surgery based on preoperative variables obtained from a surgical database of a large acute care hospital in Scotland. Our results show that adaboost has the best overall performance (AUC=0.731), however random forest (Sensitivity = 0.852, negative predictive value = 0.923) and gradient boosting model (Sensitivity = 0.875 and negative predictive value = 0.920) have the best performance at predicting severe postoperative complications based on sensitivity and negative predictive value.

Keywords: Postoperative Complications, Machine Learning, Cardiac Surgery.

1 Introduction

The 2011 National Confidential Enquiry into Patient Outcome and Death (NCEPOD) estimated that there are between 20,000-25,000 deaths among people undergoing a surgical procedure every year in the UK [1]. Approximately 80% of these deaths occur amongst a minority of 'high risk' patients, who make up approximately 10% of the overall surgical population. In addition to facing higher mortality rates, these patients also have increased risk of postoperative complications, and therefore require high levels of care and clinical resources before, during and after surgery [1].

Over the last two decades, an increasing number of hospitals have developed preoperative clinics and services [2] designed to triage patients well in advance of their surgery into 'low risk patients', suitable for day-care surgery, and 'high-risk patients', requiring additional management and admission as inpatients [3]. Data-driven risk scoring systems are now an integral component of these surgical pre-assessment clinics, and most of these generally focus specifically on predicting patients' risks of mortality [4].

According to the Society of Cardiothoracic Surgery in Great Britain and Ireland, the in-hospital mortality rate after cardiac surgery has remained low: i.e. under 3% over the past five years [5]. Although surgical mortality rates are low, complications after surgery are common, and can have an important impact on patients' quality of life [6,7].

Surgical complications can also increase hospital length of stay [8–10] and healthcare costs [11–13]. Hence, a robust and reliable predictive model for postoperative complications would prove extremely useful for managing patient flows and clinical resources in surgical care.

There are currently no validated surgical risk scoring systems available which can predict generic surgical complications and their severity [4,14]. In order to explore the feasibility of developing such a scoring system, we have previously explored various machine learning methods, such as logistic regression, random forest, naïve Bayes and bootstrap aggregated classification and regression trees at predicting severe postoperative complications in our patient population. As the percentage of patients with severe postoperative complications is relatively small compared to no or other complications, we are facing an imbalanced classification problem, which is one of the biggest challenges in prediction modeling due to its presence in many real-world classification tasks [15]. There are various methods available to approach this, including modifying existing algorithms to take into account the significance of positive examples [16] and using methods to balance datasets, such as Synthetic Minority Over-sampling Technique (SMOTE) [17].

In this paper we are presenting our results from another approach: the use of ensembles of classifiers, which has been shown to have a better performance when approaching class imbalance problems [18]. Ensembles are designed to increase the accuracy of a single classifier by training several different classifiers and combining their decisions to output a single class label [19]. The range of methods which were evaluated and compared include: random forest and ensemble methods.

This paper is structured as follows: we describe our methods in Section 2, provide our results in Section 3 and discuss the relevance of our findings in Section 4.

2 Methods

2.1 Study Setting, Cardiac Surgery Data and Categorization of Complications

Setting. This project was conducted with the Golden Jubilee National Hospital (GJNH)[1], Clydebank, Scotland. GJNH is a state-of-the art tertiary referral center, carrying out a range of major surgical procedures (general, cardiac, orthopedic and thoracic surgery) with a commitment to reducing patient national waiting times across the National Health Service (NHS) in Scotland, while striving to deliver the highest quality of care. The hospital has 15 operating theatres. In 2016/17 GJNH carried out a total of 40,929 inpatients, day cases and diagnostic examinations.

Study Ethics & Data. This study was approved by our Institution's Research and Development Review Board and classified as an anonymized data study covered by Caldicott status. Data about cardiac procedures were obtained from a clinical audit database

[1] https://www.nhsgoldenjubilee.co.uk/

called the Cardiac, Cardiology and Thoracic Health Information system (CaTHI). The database consists of cardiac, cardiology and thoracic patients' diagnostics, surgical procedures and discharge information. All admissions in cardiac surgery between 1st April 2012 and 31st March 2016 were recorded in the CATHI database, adding up to a total of n=3838 admissions. All patients reported in the CaTHI database received a treatment. In the analysis, only patients undergoing coronary artery bypass graft (CABG), valve and combined CABG and valve surgery were included in the study, the final study sample being n=3700 clinical records.

Being a clinical audit database, most variables in the CaTHI database were consistently recorded. In cases where categorical variables had missing data, the blank fields were coded as "Unknown". The variables with "Unknown" entries included renal impairment (43.38% unknown), rhythm (7.97%), smoking status (36.24%), and left main stem disease (48.76%). If a numerical variable was not recorded consistently, the variable was excluded from the analysis. The only variable excluded for that reason was preoperative hemoglobin level.

Therefore, the final dataset used for our analysis consists of 25 preoperative variables[2], including patient characteristics, preoperative variables about patients' cardiac status and comorbidities, as well as other surgical variables. Three of the variables (age, preoperative serum creatinine and body mass index (BMI)) are numerical variables, the rest of them are categorical.

Categorization of Complications. With the assistance of a panel of consultant cardiac anesthetists and surgeons in GJNH, we categorized complications reported in the CaTHI database into four discreet categories (no/mild/moderate/severe) based on their impact on hospital length of stay, patients' quality of life and cost of care. The categorization was subsequently cross-referenced with findings from a literature review we have conducted in relation to risk scoring of perioperative complications. The categorization task resulted in 3 categories of complications (mild/moderate/severe), including: 17 types of mild complications, 42 moderate complications and 19 severe complications.

2.2 Model Development

In this study, we have focused on developing a predictive model for solving a binary classification task: i.e. whether or not a patient is likely to have a severe postoperative

[2] Variables include: age, sex, diabetes, body mass index, smoking, neurological dysfunction, congestive cardiac failure, previous myocardial infarction, active endocarditis, hypertension history, New York Heart Association grade, angina status, rhythm of the heart, left ventricular function, left main stem disease, extra-cardiac arteriopathy, pulmonary disease, creatinine levels, renal function, surgical priority, critical preoperative state, surgical procedure, previous operations, previous percutaneous coronary intervention.

complication ('yes' or 'no or other') [3]. The reason why we chose to focus on predicting sever complications in the first instance is due to the fact that these have the most detrimental impact to patients and on the use of clinical resources (e.g. such as requiring additional procedures to manage the complication or increasing hospital length of stay).

As this is an imbalanced classification problem involving both categorical and numerical variables, we used machine learning methods appropriate for this kind of data analysis: *random forest*, *adaboost*, *gradient boosting model* and *two stacked models*. All analysis was conducted with statistical package R version 3.5.0.

Random Forest, Adaboost and Gradient Boosting. The random forest, adaboost and gradient boosting models were developed using *k*-fold cross-validation, where the training data (n=2479 records) was randomly partitioned into *k* sub-sets of approximately equal sizes. At each *k* iteration one of the folds is chosen as the test set and the remaining *k-1* are used for the training.

This method often results in a less biased and less optimistic estimate of the model than other methods. In this study we use 5-fold cross-validation, as is generally recommended in the literature [20].

For random forest, the package 'randomForest' version 4.6-14 [21] was used with the number of trees set at n=200. The adaboost model was developed using the package 'fastAdaboost' version 1.0.0 [22], which implements Freund and Schapire's Adaboost.M1 algorithm [23], and for which we conducted n=40 iterations. For the gradient boosting model, the package 'gbm' version 2.1.5 [24] was used, which uses the Friedman's gradient boosting algorithm [25]. The number of trees was chosen to be n=1000 and the shrinkage parameter as 0.01. For these three models, we evaluated the performance using a separate set of testing data (n=1221 records).

Stacked Models. The appropriate base learners for our data that were included in our stacked models were *generalized linear model* [26], *random forest* [27], *naïve Bayes* [28] and *bootstrap aggregated classification and regression trees (Bagging CART)* [29]. We firstly generated k-fold cross-validated predicted values from the base learners to generate the training data for the metalearner algorithm. The training set (n=1850 records) was used to develop our base learners. Then a validation set (n=925 records) was used to create the level one dataset. The base learners and the ensemble were then evaluated using the testing dataset (n=925 records). In this study we compared two different metalearner algorithms: random forest and generalized linear model. All analysis for the stacked models was done using the package 'caret' version 6.0-81 [30].

[3] Severe complications in this study include: Acute renal failure, deep sternal wound infection, septicemia, transient stroke, tracheostomy, cardiac arrest, permanent stroke, severe heart failure, adult respiratory distress syndrome, multi-organ failure, mesenteric infarction, required laparotomy, severe pulmonary edema, left ventricular wall dissection, hepatic failure, reopening requiring coronary artery bypass graft, paraparesis, and amputation.

2.3 Model Evaluation and Performance Measures

The models were evaluated based on the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity (a.k.a. recall), specificity (a.k.a. true negative rate), and positive (PPV) and negative predictive value (NPV). As this is an imbalanced classification problem, where the prevalence of severe postoperative complications is small compared to *'no or other'* complications, using these performance measures help us avoid the accuracy paradox [31].

As the aim of this study is to predict severe complications, we are aiming for the highest sensitivity and negative predictive value as possible. This is to ensure that the model recognizes as many patients with severe complications as possible *(i.e. sensitivity)* and in case of negative testing: to ensure that the probability that the patient actually does not have a severe complication is high *(i.e. negative predictive value)*.

3 Results

3.1 Population Characteristics

In our study sample of n=3700 clinical records and using the classification of complications described earlier in section 2.1, 48.65% of the patients had a recorded postoperative complications. Of these: 7.05% had mild complications, 36.65% moderate complications, and 4.95% severe complications after cardiac surgery.
As the prevalence for severe complications in our patient population is 4.95%, this is a highly imbalanced classification task.

Of all patients, 59.65% had a CABG, 26.49% had a valve surgery, and 13.86% had a combined CABG and valve surgery. The mean age was 66.7, with a median of 68 years. The majority of the patients were men (73.22%). Overall, 26.51% of the patients had diabetes. Based on body mass index, 42.46% of the patients were obese, 40.22% were overweight, 16.47% had a normal weight and 0.85% were underweight. Slightly less than a quarter of the patients (22.71%) had never smoked, 11.70% were current smokers, 29.35% were ex-smokers and for 36.24% of the patients the smoking status was unknown. The patient characteristics for patients with severe and no or other complications can be found from Table 1.

Table 1. Patient characteristics for patients with 'severe' and 'no or other' complications. For numerical variables: median, mean and standard deviation are provided, for categorical variables: frequencies and percentages are provided.

Variable	Severe N=183	No or Other N=3517
Age (median, mean ± SD)	71.5, 69.7 ± 11.1	68, 66.5 ± 10.7
Sex: Female (%)	71 (38.80)	920 (26.16)
Diabetes (%)	65 (35.52)	916 (26.04)
BMI (median, mean ± SD)	29.7, 29.9 ± 5.3	29.0, 29.5 ± 5.1
Smoking Status: Ex-smoker (%)	63 (34.43)	1023 (29.09)

Variable	Severe N=183	No or Other N=3517
Current smoker	17 (9.29)	416 (11.83)
Unknown	66 (36.07)	1275 (36.25)
Neurological Dysfunction (%)	5 (2.73)	74 (2.10)
Congestive Cardiac Failure: At admission (%)	21 (11.48)	61 (1.73)
Past	19 (10.38)	188 (5.35)
Previous MI (%)	81 (44.26)	1276 (36.28)
Active Endocarditis (%)	7 (3.83)	20 (0.57)
Hypertension History (%)	150 (81.97)	2556 (72.68)
NYHA Grade: II (%)	80 (43.72)	1841 (52.35)
III	69 (37.70)	941 (26.76)
IV	13 (7.10)	84 (2.39)
Angina Status: I (%)	18 (9.84)	483 (13.73)
II	66 (36.07)	1333 (37.90)
III	34 (18.58)	572 (16.26)
IV	8 (4.37)	173 (4.92)
Rhythm of the Heart: Abnormal (%)	141 (77.05)	287 (8.16)
Unknown	15 (8.20)	280 (7.96)
LV Function: Moderate (%)	37 (20.22)	556 (15.81)
Poor	16 (8.74)	86 (2.45)
Left Main Stem Disease (%)	23 (12.57)	451 (12.82)
Unknown	88 (48.09)	1716 (48.79)
Extracardiac Arteriopathy (%)	34 (18.58)	457 (12.99)
Pulmonary Disease (%)	47 (25.68)	651 (18.51)
Creatinine level (median, mean ± SD)	90.0, 101.0 ± 64.5	84.0, 91.3 ± 56.1
Renal Impairment: Moderate (%)	41 (22.40)	699 (19.87)
Severe	14 (7.65)	172 (4.89)
Unknown	92 (50.27)	1513 (43.02)
Surgical Priority: Emergency (%)	5 (2.73)	21 (0.60)
Prioritised	9 (4.92)	268 (7.62)
Urgent	31 (16.94)	497 (14.13)
Critical Preoperative State (%)	7 (3.83)	35 (1.00)
Surgical Procedure: CABG (%)	82 (44.81)	2125 (60.42)
Valve	62 (33.88)	918 (26.10)
Valve and CABG	40 (21.86)	473 (13.45)
Previous Cardiac Surgery (%)	164 (89.62)	73 (2.08)
Previous PCI (%)	36 (19.67)	445 (12.65)

3.2 Performance of the Models

Table 2 shows that in terms of AUC, adaboost outperforms random forest, gradient boosting and the stacked models with an AUC of 0.731. However, as our end goal is to

develop a clinical decision support system predicting severe postoperative complications, our aim is to have the highest possible sensitivity and negative predictive value. Based on that, the GBM has the highest sensitivity of 0.875, meaning that the model recognizes patients with severe complications 87.5% of the time. The GBM also has a very high negative predictive value of 0.920, which means that if the test is negative, the probability that the patient actually does not have a severe complication is 92.0%.

Table 2. Area under the curve (AUC), sensitivity, specificity, positive (PPV) and negative predictive value (NPV) for the models.

Algorithm	AUC	Sensitivity	Specificity	PPV	NPV
Random Forest	0.724	0.852	0.462	0.017	0.923
Adaboost	0.731	0.738	0.629	0.021	0.905
Gradient Boosting	0.718	0.875	0.465	0.014	0.920
Stacked with RF	0.648	0.321	0.944	0.044	0.721
Stacked with GLM	0.655	0.643	0.639	0.035	0.897

Surprisingly, both stacked models had a considerably worse performance in terms of AUC compared to the other models. In addition, the stacked model with RF has a very low sensitivity and very high specificity, which would not be useful in clinical applications.

As the random forest and gradient boosting models have the highest sensitivities and negative predictive values, we further investigated these two models. To assess which variables are the most important, for random forest we calculated at the Gini importance measure and for gradient boosting model we calculated the relative influence (Table 3).

Both of these models show ─ with some differences in ordering ─ that preoperative creatinine, BMI, age, angina status and smoking are the most important variables when predicting severe complications. These results are also supported by findings from the literature: elderly patients are at a greater risk of postoperative complications, especially for bleeding, infections, neurologic and renal problems [32].

Table 3. The importance measures for the top five variables of Gradient Boosting Model (GBM) and Random Forest (RF).

Variable	GBM (relative influence)	RF (Gini importance)
Pre.Op. Creatinine	17.93	31.89
BMI	16.41	35.24
Age	10.25	28.90
Angina Status	6.90	13.27
Smoking	6.57	12.04

Patients with a higher BMI have increased risk of wound infection, blood loss and acute kidney injury [33]. Angina status is shown to be a significant predictor of long-

term mortality [34]. Persistent smokers have a higher incidence of pulmonary complications [35] and also slower wound healing following CABG surgery [36].

4 Discussion

Our study found that postoperative complications are common (48.65% in our study population) and the most severe of these ─ although less frequent at 4.95% ─ can have a significant impact on episodes of care and use of clinical resources as well as being potentially devastating for patients' quality of life after surgery. It is therefore essential that adequate systems are developed within clinical care in order to better plan and mitigate these instances of severe perioperative complications.

Our findings from the literature identified five cardiac preoperative risk prediction models commonly used in clinical practice. These include: logistic European System for Cardiac Operative Risk Evaluation (EuroSCORE) [37], EuroSCORE II [38], the Initial Parsonnet Score [39], the Society of Thoracic Surgeons (STS) score [40,41] and the Cleveland Clinic Score [42]. The first three were developed to predict 30-day mortality, and the latter two were developed to predict mortality as well as some complications. All of these models were developed using logistic regression. In spite of these scores being initially developed to predominantly predict postoperative mortality, some studies have been carried out to assess the use of these scoring systems to predict combinations of postoperative complications [32,43–50].

Looking at the AUC, our adaboost model outperforms all aforementioned studies, apart from Parsonnet score in one study [43] and STS score another study [50], where these scores have a similar performance with the adaboost model (AUC=0.73). Our random forest model has a similar performance to EuroSCORE and EuroSCORE II in one study (AUC=0.72 for both) [50]. Even though our GBM model has the lowest performance out of these three in our study, it still outperforms the commonly used risk models in most aforementioned studies, apart from EuroSCORE, EuroSCORE II, STS [50] and Initial Parsonnet [43] in two studies.

Performance Measures. Even though the adaboost model has the highest AUC, the performance of sensitivity and negative predictive value are the most important for the purpose of developing a decision support application for severe complications. A model with a high specificity can be used to rule out patients who do not need specific treatment [51]. However, our aim is to develop a model which can identify which patients are more likely to develop severe postoperative complications, in order to improve care planning, management and monitoring. Having a higher negative predictive value, meaning the patient probably does not have the disease when the test is negative, reassures the provider of the treatment to do no harm.

Some of the previously mentioned papers evaluating the commonly used preoperative risk tools predicting complications have similar results based on AUC as our models. However, these studies have not reported other performance measures such as sensitivity, specificity, PPV and NPV.

Current Challenges in Predicting Postoperative Complications. At present a major obstacle in predicting postoperative complications is that there is currently no single nomenclature of surgical complications; unlike for clinical diagnosis (i.e. the International Statistical Classification of Diseases, ICD-10[4]). Due to that, when comparing our results with the literature, all of the aforementioned studies have a different definition for "morbidity", which includes a different set of combined complications. The reporting of different complication outcomes in the scientific literature therefore prevents the objective comparison of the performance of these predictive risk models.

It is also worth mentioning that common risk scoring systems were developed using logistic regression. Logistic regression based models have demonstrated very good performance when applied at the population level [37,38], i.e.: their prediction accuracy generally performs well when applied to broad group or categories of patients. However, the prediction performance of these models at the 'individual' level is in fact far less satisfactory [53].

Conclusion and Future Work. In this study, we have highlighted how the use of machine learning techniques could be applied to the problem of predicting postoperative complications and compared the performance of several approaches.

Through our analysis we found two machine learning models suitable for predicting severe postoperative complications: random forest and gradient boosting model based on sensitivity (0.852 and 0.875, respectively) and negative predictive value (0.923 and 0.920, respectively). Either of these models could help a clinician to identify patients who are at risk of having severe postoperative complications in order to allocate resources or avoid high-risk treatments. In order to develop a usable clinical decision support system that relies on the models developed in this study, a further validation study needs to be undertaken.

References

1. Findlay GP, Goodwin APL, Protopapa K, Smith NCE, Mason M. *Knowing the Risk: A Review of the Peri-Operative Care of Surgical Patients*. London, 2011.

2. Bouamrane M-M, Mair FS. Implementation of an integrated preoperative care pathway and regional electronic clinical portal for preoperative assessment. *BMC Medical Informatics and Decision Making* 2014; **14**.

3. Bouamrane M-M, Mair FS. A study of clinical and information management processes in the surgical pre-assessment clinic. *BMC Medical Informatics and Decision Making* 2014; **14**.

4. Moonesinghe SR, Mythen MG, Das P, Rowan KM, Grocott MP. Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: qualitative systematic review. *Anesthesiology* 2013; **119**: 959–81.

5. SCTS. Blue Book Online. *Blue Book Online*, 2016. http://bluebook.scts.org/.

[4] http://apps.who.int/classifications/icd10/browse/2010/en

6. Maillard J, Elia N, Haller CS, Delhumeau C, Walder B. Preoperative and early postoperative quality of life after major surgery - A prospective observational study. *Health and Quality of Life Outcomes* 2015; **13**: 12.

7. Pinto A, Faiz O, Davis R, Almoudaris A, Vincent C. Surgical complications and their impact on patients' psychosocial well-being: A systematic review and meta-analysis. *BMJ Open* 2016; **6**.

8. Al-Sarraf N, Thalib L, Hughes A et al. The effect of preoperative renal dysfunction with or without dialysis on early postoperative outcome following cardiac surgery. *International Journal of Surgery* 2011; **9**: 183–7.

9. Knapik P, Ciesla D, Borowik D, Czempik P, Knapik T. Prolonged ventilation post cardiac surgery - tips and pitfalls of the prediction game. *Journal of Cardiothoracic Surgery* 2011; **6**: 158.

10. Ruel M, Chan V, Boodhwani M et al. How detrimental is reexploration for bleeding after cardiac surgery? *The Journal of Thoracic and Cardiovascular Surgery* 2018; **154**: 927–35.

11. Eappen S, Lane BH, Rosenberg B et al. An Electronic Reprint Relationship Between Occurrence of Surgical Complications and Hospital Finances Relationship Between Occurrence of Surgical Complications and Hospital Finances. *JAMA* 2013; **309**: 1599–606.

12. Wang FD, Chang CH. Risk factors of deep sternal wound infections in coronary artery bypass graft surgery. *The Journal of cardiovascular surgery* 2000; **41**: 709–13.

13. Salehi Omran A, Karimi A, Ahmadi SH et al. Superficial and deep sternal wound infection after more than 9000 coronary artery bypass graft (CABG): incidence, risk factors and mortality. *BMC infectious diseases* 2007; **7**: 112.

14. Barnett S, Moonesinghe SR. Clinical risk scores to guide perioperative management. *Postgraduate Medical Journal* 2011; **87**: 535–41.

15. Yang Z, Tang WH, Shintemirov A, Wu QH. Association Rule Mining-Based Dissolved Gas Analysis for Fault Diagnosis of Power Transformers. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)* 2009; **39**: 597–610.

16. Zadrozny B, Elkan C. Learning and making decisions when costs and probabilities are both unknown. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2001: 204–13.

17. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research* 2002; **16**.

18. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transaction on Systems, Man, and Cybernetics - Part C: Applications and Reviews* 2012; **42**: 463–84.

19. Polikar R. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* 2006; **6**: 21–45.

20. Bischl B, Mersmann O, Trautmann H, Weihs C. Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary computation* 2012; **20**: 249–75.

21. Breiman L, Cutler A, Liaw A, Wiener M. Package “randomForest.” 2018.

22. Chatterjee S. Package “fastAdaboost.” 2016.

23. Freund Y, Schapire RE. Experiments with a new boosting algorithm. *The Thirteenth International Conference on Machine Learning*. 1996: 148–56.

24. Greenwell B, Boehmke B, Cunningham J. Package “gbm.” 2019.

25. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 2001; **29**: 1189–232.

26. Walker SH, Duncan DB. Estimation of the Probability of an Event as a Function of Several Independent Variables. *Biometrika* 1967; **54**: 167–79.

27. Ho TK. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1998; **20**: 832–44.

28. Zhang H. The Optimality of Naive Bayes. *In FLAIRS2004 Conference*. 2004.

29. Breiman L. Bagging Predictors. *Machine Learning* 1996; **24**: 123–40.

30. Kuhn M. Package “caret.” 2018. https://cran.r-project.org/web/packages/caret/caret.pdf.

31. Valverde-Albacete FJ, Pelaez-Moreno C. 100% classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox. *PLoS ONE*2014; doi https://doi.org/10.1371/journal.pone.0084217.

32. Wang TK, Li AY, Ramanathan T, Stewart RA, Gamble G, White HD. Comparison of four risk scores for comtemporary isolated coronary artery bypass grafting. *Heart, lung and circulation* 2014; **23**: 469–74.

33. Reis C, Barbiero SM, Ribas L. The effect of the body mass index on postoperative complications of coronary artery bypass grafting in elderly. *Revista brasileira de cirurgia cardiovascular* 2008; **23**: 524–9.

34. Kaul P, Naylor CD, Armstrong PW, Mark DB, Theroux P, Dagenais GR. Assessment of activity status and survival according to the Canadian Cardiovascular Society angina classification. *The Canadian journal of cardiology* 2009; **25**: e225-31.

35. Ji Q, Zhao H, Mei Y, Shi Y, Ma R, Ding W. Impact of smoking on early clinical outcomes in patients undergoing coronary artery bypass grafting surgery. *Journal of Cardiothoracic Surgery* 2015; **10**.

36. Sharif-Kashani B, Shahabi P, Mandegar M-H et al. Smoking and wound complications after coronary artery bypass grafting. *Journal of Surgical Research* 2016; **200**: 743–8.

37. Roques F, Michel P, Goldstone AR, Nashef SAM. The logistic EuroSCORE. *European Heart Journal* 2003; **24**: 1–2.

38. Nashef SA, Roques F, Sharples LD et al. EuroSCORE II. *European Journal of Cardio-Thoracic Surgery* 2012; **41**: 734–44.

39. Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. *Circulation* 1989; **79**: I3-12.

40. Shroyer AL, Coombs LP, Peterson ED et al. The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models. *The annals of thoracic surgery* 2003; **75**: 1856–64.

41. Shahian DM, O’Brien SM, Fillardo G et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1-- coronary artery bypass grafting surgery. *The*

annals of thoracic surgery 2009; **88**: S2-22.

42. Higgins TL, Estafanous FG, Loop FD, Beck GJ, Blum JM, Paranandi L. Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients. *JAMA* 1992; **267**: 2344–8.

43. Dupuis J-Y, Wang F, Nathan H, Lam M, Grimes S, Bourke M. The cardiac anesthesia risk evaluation score: a clinically useful predictor of mortality and morbidity after cardiac surgery. *Anesthesiology* 2001; **94**: 194–204.

44. Gabrielle F, Roques F, Michel P et al. Is the Parsonnet's score a good predictive score of mortality in adult cardiac surgery: assessment by a French multicentre study. *European journal of cardio-thoracic surgery* 1997; **11**: 406–14.

45. Geissler HJ, Holzl P, Marohl S et al. Risk stratification in heart surgery: comparison of six score systems. *European journal of cardio-thoracic surgery* 2000; **17**: 400–6.

46. Hirose H, Inaba H, Noguchi C et al. EuroSCORE predicts postoperative mortality, certain morbidities, and recovery time. *Interactive CardioVascular and Thoracic Surgery* 2009; **9**: 613–7.

47. Pitkänen O, Niskanen M, Rehnberg S, Hippelainen M, Hynynen M. Intra-institutional prediction of outcome after cardiac surgery: comparison between a locally derived model and the EuroSCORE. *European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery* 2000; **18**: 703–10.

48. Scolletta S, Giomarelli P, Cevenini G, Biagioli B. Estimation of morbidity risk factors in intensive care unit: a Bayesian discriminant approach: 028. *European journal of anaesthesiology* 2004; **21**: 14.

49. Syed AU, Fawzy H, Farag A, Nemlander A. Predictive value of EuroSCORE and Parsonnet scoring in Saudi population. *Heart, lung and circulation* 2004; **13**: 384–8.

50. Wang TKM, Harmos S, Gamble GD, Ramanathan T, Ruygrok PN. Performance of contemporary surgical risk scores for mitral valve surgery. *Journal of cardiac surgery* 2017; **32**: 172–6.

51. Lutkenhoner B, Basel T. Predictive modeling for diagnostic tests with high specificity, but low sensitivity: a study of the glycerol test in patients with suspected meniere's disease. *PLoS ONE* 2013; **8**.

52. Harty J. Prevention and Management of Acute Kidney Injury. *Ulster Medical Journal* 2014; **83**: 149–57.

53. Alaa AM, Yoon J, Hu S, van der Schaar M. Individualized Risk Prognosis for Critical Care Patients: A Multi-task Gaussian Process Model. 2017: 1–10.