

# 1                   **Random effects adjustment in machine learning models for cardiac** 2                   **surgery risk prediction: a benchmarking study**

3   Tim Dong MSc<sup>1</sup>, Shubhra Sinha MD<sup>1</sup>, Daniel P Fudulu PhD<sup>1</sup>, Jeremy Chan MD<sup>1</sup>, Ben Zhai PhD<sup>3</sup>, Pradeep  
4   Narayan FRCS(CTh)<sup>2</sup>, Massimo Caputo MD<sup>1</sup>, Andy Judge PhD<sup>1</sup>, Arnaldo Dimagli MD<sup>1</sup>, Umberto  
5   Benedetto PhD<sup>1</sup> and Gianni D. Angelini MD<sup>1</sup>.

6  
7  
8   <sup>1</sup>Bristol Heart Institute, Translational Health Sciences, University of Bristol

9   <sup>2</sup>Department of Cardiac Surgery, Rabindranath Tagore International Institute of Cardiac Sciences, India

10   <sup>3</sup>Department of Computing Science, Northumbria University

11  
12  
13  
14  
15   Mr Tim Dong,

16   British Heart Foundation DSC Project Data Science Lead.

17   Bristol Heart Institute,

18   Bristol, BS2 8HW

19   Phone number: +447564168791

20   Email: [qd18830@bristol.ac.uk](mailto:qd18830@bristol.ac.uk)

28 **ABSTRACT**

29 **Objectives** There is an ongoing debate over whether a procedural specific (e.g. Society of Thoracic  
30 Surgeons (STS)) or universal model (e.g. EuroSCORE II (ES II)) should be used for patient selection in  
31 cardiac surgery. Recently, we showed that ES II suffers from severe performance drift across several  
32 important metrics and that ML approaches such as Xgboost and Random Forest are substantially more  
33 resistant to dataset drift. With the growing interest in big data and its leverage through the use of ML  
34 approaches that are not limited by linear statistical assumptions, the number of clinical variables can  
35 theoretically increase exponentially. In addition, the variations and residual confounding that historically  
36 hindered the usefulness of cardiac risk stratification scores can potentially be taken into account. Here, we  
37 assess these possibilities on a large United Kingdom (UK) database.

38 **Methods:** A retrospective analysis of prospectively routinely gathered data on adult patients undergoing  
39 cardiac surgery in the UK between 2012-2019. We temporally split the data 70:30 into a training and  
40 validation subset. Two sets of seven ML mortality prediction models, with and without variable selection  
41 were assessed for consensus Clinical Effective Metric (CEM) overall performance and performance within  
42 each of CEM's constituent metrics. Confounding and potential causal relationships between covariates and  
43 outcomes were evaluated using bayesian network analysis.

44 **Results:** A total of 227,087 adults underwent cardiac surgery during the study period with a mortality rate  
45 of 2.76%. For non-variable selected (NVS) risk scores with 102 variables, Xgboost with adjustment for  
46 hospital variation was superior to the Xgboost without adjustment ( $p < 2e-16$ ). Both NVS and the 18  
47 variables selected (VS) Xgboost with adjustment for hospital variation risk scores were superior to the  
48 Xgboost (ES II 18 variables) model ( $p < 6.3e-15$ ), with NVS Xgboost with adjustment for hospital variation  
49 having the best performance, followed by the VS Xgboost with adjustment for hospital variation (CEM  
50 Difference: 0.0150 and 0.0023, respectively).

51 **Conclusions:** We have identified an ML adjusted risk score comprising 102 variables that increases risk  
52 stratification performance on hold out dataset, removing the need to perform variable selection and

53 reduction. This paves the way for further research that utilises this new set of variables with hospital-based  
54 adjustments for the safer selection of patients undergoing cardiac surgery.

55 **Keywords:** cardiac surgery; artificial intelligence; risk prediction; machine learning; operative mortality;  
56 confounding analysis; causal analysis; performance analysis; national dataset

57

58

59

50

51

52

53

54

55

56

57

58

59

70

71

72

73

74

75

76

## 77 Introduction

78 The importance of Machine Learning (ML), a branch of Artificial intelligence (AI) has recently been  
79 highlighted as a potential alternative to mortality risk scores for cardiac surgical procedures, such as Society  
30 of Thoracic Surgeons (STS),[1] and EuroSCORE II (ES II) [2] which are prone to miscalibration overtime  
31 and poor generalisation across datasets.[1,3] In particular, ES II, which is based on logistic regression using  
32 18 items of patient information, has been shown by numerous studies to display poor discrimination and  
33 calibration across datasets with differing characteristics, including but not limited to age,[4] ethnicity[5] and  
34 procedures groups.[6–10] Furthermore, ES II suffers from severe performance drift across several metrics  
35 including but not limited to discrimination, calibration, clinical utility and overall accuracy. ML approaches  
36 such as Xgboost and Random Forest are substantially more resistant to performance drifts that arise as a  
37 consequence of dataset drift.

38 ML has been shown to be superior to conventional scoring systems with the magnitude and clinical  
39 influence of such improvements demonstrated.[2] The ability to counter-performance drift due to temporal  
30 changes in the prevalence of risk factors has also been evaluated across multiple centres and has been shown  
31 to be superior to universal scores such as the ES II. However, the confounding effects of variables not  
32 included in ES II for its constituent procedures have not been taken into account and may confuse  
33 correlation with causation of the outcome,[11] and may also limit the full potential of the risk stratification  
34 scores. The influence of these “hidden” variables on the performances of ML scores have yet to be fully  
35 elucidated.

36 Parsimonious models can result in improved prediction by preventing overfitting in scenarios where  
37 the number of events to variables are low, e.g. small sample size and high dimensional datasets. However,  
38 The No Free Lunch Theorem states that all optimisation algorithms perform equally well when their  
39 performance is averaged across all possible problems,[25] and suggests that different ML models will have a  
30 different set of optimal prediction variables for any given task or dataset. This makes it difficult to provide a

31 fair comparison of models based on equal events per variable (EPV). Keeping the EPV constant would  
32 require different sample size of data being compared due to differences in number of variables per model.  
33 This would lead to unfairness due to sample size disparity. An alternative solution to create a comparative  
34 set of parsimonious models is to keep EPV constant by using the top  $N$  most important variables across all  
35 models, thereby also fixing the sample size across all model comparisons.

36 Since cardiac surgery mortality events are low and typically in the order of 3%, the events per  
37 variable (EPV) will require careful judgement of the right balance of variables to use for prediction  
38 modelling.[12] In small datasets, the cardiac surgery risk predictions are more likely to be biased and to  
39 have high error rates, especially when larger numbers of variables are used.[13] The commonly used ES II is  
40 limited to the use of 18 variables partly for this reason. However, for larger multi-centre National Adult  
41 Cardiac Surgery Audit (NACSA) datasets, the number of events will be larger relative to smaller datasets,  
42 theoretically supporting the use of larger number of variables. One other reason that ES II does not use a  
43 larger number of variables is due to restrictions in logistic regression that variables should be absent of  
44 multi-collinearity i.e. predictors should be independent of each other. This assumption becomes more  
45 difficult to meet as variable numbers increases. However, ML is not affected by this limitation and is more  
46 suited to modelling complex non-linear relationships among variables. We therefore hypothesised whether  
47 ML could be applied to enable increased performance when a larger set of variables are used.

48 One other issue of using a larger multi-centre national dataset such as NACSA is that there may exist  
49 systematic differences in relationships of variables across different centres, i.e. calibration drift across  
50 geographic regions. For example, each hospital could have different cardiac surgery protocols or different  
51 suppliers for surgical equipments and devices. In addition, patients from different centres are likely to have  
52 different levels of deprivation and social demographic profiles that could result in regional differences in  
53 cardiac surgery risk. Therefore, we hypothesise that taking hospital variation into account could improve the  
54 performance of cardiac surgery risk prediction.

55 We, therefore, trained and evaluated two sets of 7 supervised ML models based on various  
56 combinations of a large set of 60 clinical variables with and without hospital location variable(s) to (1)  
57 determine the best ML model in terms of overall accuracy, discrimination, calibration and clinical

28 effectiveness, (2) use variable importance to select and build a parsimonious version of the set of 7 models  
29 from (1) and to compare performances within and across parsimonious and non-parsimonious sets; (3)  
30 analyse causal and non-causal relationships between newly proposed variables and the outcome.

31

## 32 Related works

33 Machine learning (ML) techniques have drawn interest as potential substitutes for existing scoring systems  
34 for predicting the mortality risk of cardiac surgery. ML models have been demonstrated to perform better  
35 than traditional methods, offering predictions that are more accurate with potential treatment  
36 optimization applications. However, prior studies of cardiac surgery risk prediction modelling using ML  
37 have mostly concentrated on applying ML algorithms without taking into account the potential variability of  
38 patients across various hospital locations.

39         Incorporating random effects into machine learning models for risk prediction applications has  
40 drawn attention in recent years. By incorporating random effects, this method, sometimes referred to as  
41 mixed effects machine learning, may take into account the heterogeneity introduced by various hospital  
42 locations. The model can capture the systematic variations in the interactions between variables across  
43 different centres through the incorporation of random effects, which enhances the risk prediction model's  
44 overall performance and calibration.

45         While one survival analysis study using ML on cardiovascular related risk factors highlighted the  
46 importance of using the hospital index as a random effect to account for clustering within the hospital,[14]  
47 this random effects was only evaluated for the traditional Cox proportional-hazards regression as part of the  
48 sub-analysis without similar adjustment in the Random Forest model. While there are few studies specific to  
49 mixed machine learning in cardiac surgery, one study in the related field of cardiovascular imaging applied  
50 linear mixed effects model as a post-processing step following convolutional neural network (CNN) to  
51 compare across different coefficient of variation across decision made by human and the CNN.[15]  
52 However, the no mixed effects machine learning or deep learning model was used.

53 In other domains such as neurology[16] mixed effects linear regression have also been used to  
54 account for random variations across models compared. However, such studies also apply mixed effects  
55 separate to the main models analysed rather than incorporating effects with the machine learning model.  
56 Unlike previously mentioned studies which applied linear mixed effect models to compare machine learning  
57 models, another study in precision oncology used random effects within a linear mixed model to select  
58 features while adjusting for patient variations.[17] In a different study on plant sciences utilising a random  
59 forest, similar linear mixed random effects selection of features before machine learning modelling was  
60 used.[18]

61 We have previously evaluated the calibration changes in machine learning base models across the  
62 1996-2011 and 2011-2017 for the EuroSCORE I variables and shown that both LR and random forest  
63 models were associated with good discrimination ability but substantial miscalibration.[19] We followed  
64 this by a development of a suit of performance metrics for evaluating clinical machine learning models.[20]  
65 In a separate study, we have developed an approach that compared calibration changes, variable importance  
66 drift, performance drift and actual dataset drift of the base models using EuroSCORE II variables across the  
67 years 2017-2019.[21] With respect to changes in techniques, Dataset drift was observed across the Holdout  
68 time periods for Weight of intervention of EuroSCORE II. Sharp dataset drifts were observed for the Single  
69 non-CABG and 3 procedures category between 2018-12 to 2019-02. In a separate study, we found that  
70 machine learning models could be ensembled to combine EuroSCORE I and EuroSCORE II variables and  
71 data from different time periods to improve performance through Xgboost homogeneous ensembles.[22]

72 While prior research has emphasised the need for mixed effects in machine learning studies and  
73 related techniques in prediction, little research has specifically compared the different ways of  
74 encoding random effects to account for hospital site heterogeneity in ML models. By using a mixed effects  
75 machine learning approach with random effects as hospital location, this research aims to bridge this gap.  
76 The suggested model can effectively account for regional variations in cardiac surgery operation and patient  
77 characteristics by taking into consideration the hospital-specific impacts. As a result, the model's predictions  
78 of mortality risk are more precise and reliable.

79

## 30 Materials and methods

31

32 The register-based cohort study is part of a research approved by the Health Research Authority (HRA) and  
33 Health and Care Research Wales and a waiver for patients' consent was waived (HCRW) (IRAS ID:  
34 278171). An Abbreviations and Definitions list of frequently used technical terms used in this study has  
35 been provided for the reader at the start of the Supplementary Materials.

### 36 Dataset and Patient Population

37 The study was performed using the National Adult Cardiac Surgery Audit (NACSA) dataset, which  
38 comprises data prospectively collected by National Institute for Cardiovascular Outcome Research on all  
39 cardiac procedures performed in all NHS hospital sites and some private hospitals across the UK.[19]

40 Patients undergoing cardiac surgery from 42 cardiac surgery centres between 1 Jan 2012 and 31 Mar  
41 2019 were included. Missing and erroneously inputted data in the dataset were cleaned according to the  
42 National Adult Cardiac Surgery Audit Registry Data Pre-processing recommendations;[23] details are found  
43 in the Supplementary Materials, Table S1: Handling of missing data and Supplementary Materials  
44 Treatment of Missing Data section. Missing categorical variable values were generally set to the baseline  
45 level, i.e., no risk were present, except where other specific values are more appropriate. Missing continuous  
46 variable values were imputed using the median (Hmisc R package). Detailed variable processing are shown  
47 in Table S1. Variable distributions were checked using histogram plots. Data standardization was performed  
48 by subtracting variable mean and dividing by the standard deviation values.[24]

49 The dataset was split into two cohorts: Training/Validation (n = 157196; 2012-2016) and Holdout (n  
50 = 69891; 2017-2019) as per previous studies.[20,21]

51

### 52 Baseline Statistical analysis



33 Continuous variables are compared using non-parametric Wilcoxon rank-sum test, whilst categorical  
34 variables are compared using Pearson's  $\chi^2$  tests or Fisher's exact test as appropriate. Baseline variable  
35 characteristics were assessed by pooling the top 18 most important variables identified through SHAP from  
36 each model and retaining only unique variables across all models.

37 Scikit-learn v0.23.1 and Keras v2.4.0 were used to develop the models and to evaluate their discrimination,  
38 calibration and clinical effectiveness capabilities. Statistical analyses are conducted using STATA-MP  
39 version 17 and R v4.0.2.[25] Anova Assumptions were checked using R rstatix package.

10

## 11 Variable Selection and Processing

12 A total of 245 NACSA variables were considered. 179 indication, intra-op, anatomical, rare procedures,  
13 dates, comorbidities, other outcomes and similar variables were excluded. An additional 3 intra-op variables  
14 were excluded resulting in 63 variables. Two variables (ethnicity and the total number of grafts) were  
15 excluded because they were only recorded at 1 hospital. Intra-Aortic Balloon Pump usage (IABP) and  
16 ventricular assist device used (VAD) were recoded to two levels (pre-op usage: Yes or No). Aortic valve  
17 (AV), Mitral Valve (MV), Tricuspid Valve (TV) and Pulmonary valve (PV) procedures were recoded to  
18 two levels (procedure performed: Yes or No). Pre-op sinus, Atrial Fibrillation (AF), Ventricular Fibrillation  
19 or Tachycardia, and heart block or paced rhythm were combined as a single categorical variable with levels  
20 0, 1, 2, 3, respectively. Aortic, Tricuspid, Pulmonary and Mitral valve procedures were made more general  
21 by combining individual repair types into a single repair category, resulting in three levels: 0.None; 1.Repair;  
22 2.Replacement. Further variable processing details are provided in Table S1, resulting in 61 initial sets of  
23 variables (Figure S2). A correlation analysis was conducted to determine the collinearity of variables.

24

## 25 Fixed and mixed effects dataset

26 Three different versions of datasets were generated based on whether the geographic location of the cardiac  
27 centre was 1) modelled as mixed-effects model or 2) not; 3) excluded. The number of baseline variables for  
28 1) is 61; 2) is  $60 + 42 = 102$ ; and for 3) is 60.

### 29 1) Mixed Effects modelling

30 The geographical location of the 42 centres was converted from the character format to a single vector of  
31 numeric equivalents and entered as the random effects.

### 32 2) Fixed Effects modelling

33 The geographical location of the 42 centres was hot-encoded into 42 new variables each indicating whether  
34 or not each procedure originated from that geographical location.

### 35 3) No Cardiac Centre

36 This is the baseline model with the removal of the geographical location variable.

## 37 Non-variable selected Modelling

38 Using the above datasets, seven models were developed without variable selection. Those model included  
39 Xgboost – centres,[26] Xgboost + centres (hot encoded), RF – centres,[27] RF + centres (hot encoded),  
40 Mixed Effects RF (MERF) + random effects (RE): centre, Mixed Effects Xgboost + RE: centre and  
41 GPBoost + RE: centre, where – indicates the exclusion of centre variable and + indicates inclusion of centre  
42 variable. Due to the MERF and Mixed-effects Xgboost requiring the outcome to be in a continuous format,  
43 the outcome variable for these two models was transformed into probabilities based on the corresponding  
44 training set. Xgboost – centres was considered as the baseline comparison.

## 45 Variable selected Modelling

46 Shapley global variable importance was used to identify the top 18 most important variables for each of the  
47 above models respectively.[28]  $N = 18$  was chosen in so as to enable comparison to the performance of ES II  
48 and ML models built using ES II. The identified variables was used to re-build a parsimonious version of the  
49 above set of seven models. Due to the MERF and Mixed effects Xgboost requiring the outcome to be in

50 continuous format, the outcome variable for these two models was transformed into probabilities based on  
51 the corresponding variable selected training set. Xgboost – centres was considered as the baseline  
52 comparison.

53 Further details on model development can be found in Supplementary Materials, section: Model  
54 Specification.

## 55 Hyperparameter Tuning

56 For non-variable selected models, a shuffled and stratified randomized 3-fold cross-validation (CV) search  
57 on hyperparameters was conducted for Xgboost - centres, Xgboost + centres (hot-encoded), RF - centres, RF  
58 + centres (hot-encoded). The process was repeated for the top 18 variables identified from shapley for each  
59 corresponding model. For non-variable selected models, the optimal hyperparameters for RF - centres and  
50 Xgboost - centres were used as initial hyperparameters for the Mixed-effects RF (RE: centre) and Mixed-  
51 effects Xgboost (RE: centre), respectively and the maximum number of expectation maximization (EM)  
52 iterations was set to five. For variable selected models: Mixed-effects RF (RE: centre) and Mixed-effects  
53 Xgboost (RE: centre) models, optimal hyperparameter values were identified using RF and Xgboost with  
54 shuffled and stratified randomized 3-fold cross-validation (CV) search on the top 18 variables from the  
55 corresponding non-variable selected models. Shuffled and stratified CV hyperparameter tuning was not  
56 possible for GPBoost, so a randomised 3-fold CV search on hyperparameters was conducted. In order to  
57 determine the optimal hyperparameters from the set of possible parameters, including parameters informed  
58 by previous studies,[20] 30 different combinations were randomly selected and evaluated in each fold of CV.  
59 This hyperparameter selection process was conducted for both the non-variable selected and variable  
70 selected sets of models.

71

## 72 Assessment of model performance

73 External validation was performed on the Holdout dataset (2017-2019).[29] Each model calculated the  
74 probability of surgical mortality for each patient. As per previous studies,[21,22] we applied the consensus

75 metric approach of Clinical Effectiveness Metric (CEM), using the combined geometric average results of  
76 all metrics[30]:

- 77 1. Discrimination: AUC[31], F1 score[32]
- 78 2. Calibration: 1 - ECE.[33]
- 79 3. Overall accuracy[30]: 1 - Brier score.[34]
- 30 4. Clinical utility – Net benefit Analysis[35]

31 Further details on the theory and application of each individual metric above is provided in our previous  
32 study of machine learning models using ES II variables.[20] One thousand bootstrap samples were taken for  
33 the CEM and its constituent metrics. Further details of this metric can also be found in the Supplementary  
34 Materials, section: Assessment of model performance.

35 We evaluated the following comparisons:

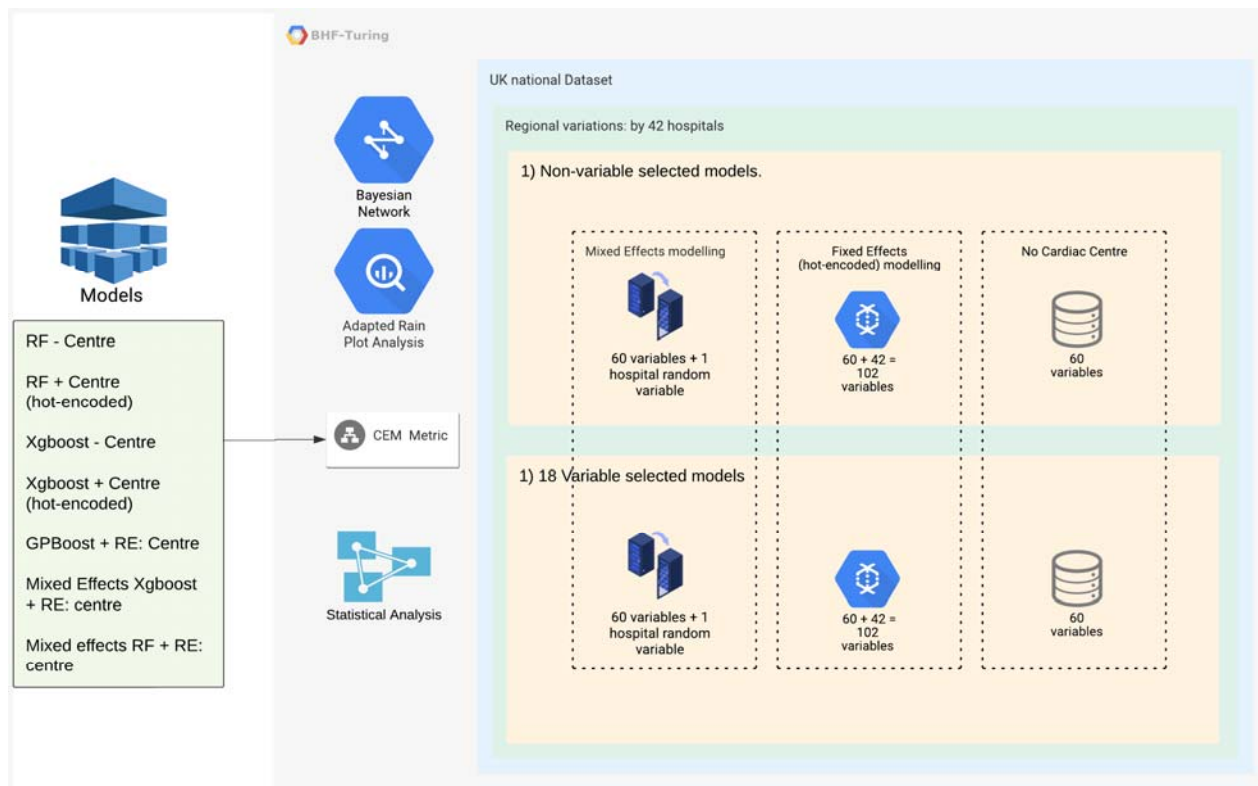
- 36 1) Non-variable selected Xgboost – centres model against all other non-variable selected models.
- 37 2) Variable selected Xgboost – centres model against all other variable selected models.
- 38 3) Highest performing model from 1) against that from 2) and Xgboost built using EuroSCORE II (ES  
39 II) variables, with the latter model as the control.[ref prj 1.2]

30 For comparisons 1) and 2), adapted Rain plots in R-3.6.2 was used to visualise constituent metrics within  
31 CEM.[36] For comparisons 1), 2) and 3), differences across models' CEMs were tested using Repeated  
32 measures One-Way Anova and Bonferroni Corrected multiple pairwise paired t-tests; this was followed by  
33 Dunnett's Correction for multiple comparisons. ANOVA assumptions for outliers were checked. Normality  
34 assumptions were checked using the Shapiro-Wilk test and histogram plots.[37] An overview of the study  
35 design is shown in Figure 1.

36

37

38 Figure 1. Design overview of the study; Non-variable and variable selected analyses are performed; CEM is used to  
 39 simultaneously assess discrimination, calibration, clinical utility, overall accuracy; analyses are performed for i)  
 40 Mixed effects modelling with 60 main variables and hospitals as a single random variable; ii) Fixed effects modelling  
 41 with 60 main variables and 42 hospitals each as a single variable and iii) Without consideration of hospital variation.



42

43

## 44 Bayesian Networks

45 A Bayesian network is produced using the training dataset to interpret 1) the relationship between the 18  
 46 most important variables and the outcome, for the highest overall performing model and;[38] 2) the  
 47 relationships between variables in 1) and the other variables used for the non-variables selected models.

48

## 49 Results

### 50 Patients characteristics

51 A total of 227,087 procedures of patients over 18 years of age from 42 UK hospitals were included in this  
 52 analysis, following the removal of 3,930 congenital cases, 1,586 transplant and mechanical support device

13 insertion cases and 3,395 procedures missing information on mortality (Table 1). There were 6,258 deaths  
14 (mortality rate of 2.76%). The primary outcome of this study was in-hospital mortality. A CONSORT flow  
15 diagram is shown in Figure S1.1. Missing rates of variables were low except for Left Ventricular Function,  
16 pulmonary artery systolic pressure (PAsys), and Number of Valves (Figure S1.2). Missing variables were  
17 backfilled using other informative variables according to NACSA dataset cleaning protocol:  
18 <https://www.nicor.org.uk/wp-content/uploads/2018/09/nacsacleaning10.3.pdf> and then imputed to improve  
19 variable quality, after which there were no missing variable values.

20

## 21 Baseline variable characteristics

22 Following data pre-processing, there were no missing values (Figure S4). Correlation analysis of all 61  
23 variables considered for the non-variable selected models showed that there was no concern for multi-  
24 collinearity (Figure S3). After pre-processing and pooling 18 most important variables from each model and  
25 retaining only unique variables, 27 variables were found to have strong evidence of being associated with  
26 outcome (Table 1,  $p < 0.001$ ).

27

## 28 Variable Importance Characteristics

29 The top 18 important variables for each of the non-variable selected models are shown in Table S12.  
30 Detailed importance scores for the variables of each model are shown for Mixed Effects Xgboost: Table S4,  
31 Figure S5, S6; MERF: Table S5, Figure S7, S8; Xgboost – Centre: Table S6, Figure S9, S10; Random  
32 Forest (RF) – Centre: Table S7, Figure S11, S12; Random Forest + Centre (hot-encoded): Table S8, Figure  
33 S13, S14; Xgboost + Centre (hot-encoded): Table S9, Figure S15, S16; GPBoost + Centre: Table S10,  
34 Figure S17 and S18. It can be seen that models of the same general type without a centre and that with a  
35 centre (hot-encoded) have a more similar ranking of important variables than the corresponding mixed-  
36 effects model. Urgency has the highest frequency of being selected as the most important predictor of  
37 mortality across models, followed by Age, Creatinine, NYHA, Previous Surgery, Pulmonary Artery Systolic  
38 Pressure and Weight.

## 39 Hyperparameter Tuning

40 The optimal hyperparameters for non-variable selected and variable selected models are shown in Table S2  
41 and S3, respectively.

## 42 Assessment of model performance

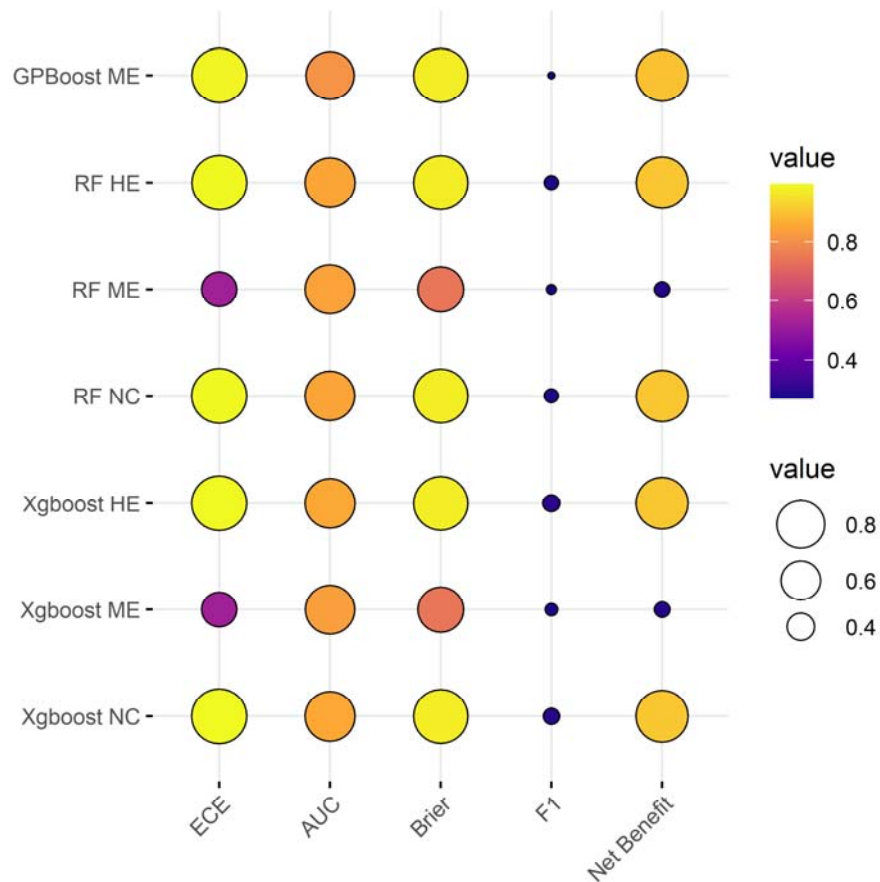
### 43 Non-variable selected model

#### 44 Adapted Rain plots

45 For non-variable selected models, it can be seen in Figure 2a that AUC performance is generally consistent  
46 across models with GPBoost having slightly lower AUC than other models. Mixed-effects RF + RE: centre  
47 and Mixed Effects Xgboost + RE: centre models performed poorly compared to other models across several  
48 metrics, namely: adjusted ECE, adjusted Brier, and net benefit. Apart from these two models, the other  
49 models performed comparably in terms of net benefit. These two models achieved higher F1 scores than  
50 GPBoost but were outperformed by all remaining models. Xgboost + Centre (hot-encoded) demonstrated the  
51 highest overall performance in terms of magnitude and ranking as shown in a detailed report of individual  
52 metric results comprising the CEM (Table 2).

53 Figure 2. a) Non variable selected models: adapted Rain plot of CEM constituent metrics by model; larger sized  
54 spheres represent higher metric performance and vice versa. ECE: 1-ECE; Brier: 1-Brier.

55



56  
57

### 58 Statistical Analyses

59 No extreme outliers were found. The CEM scores was normally distributed for all models, as assessed by  
 60 Shapiro-Wilk’s test ( $p > 0.05$ ). There was strong evidence of a difference across models  $p < 0.0001$ , except  
 61 between RF + Centre (hot-encoded) and RF - Centre (Table S11 and Figure S19). Dunnett’s test showed that  
 62 there was strong evidence that Xgboost + Centre (hot-encoded) was superior to the Xgboost - Centre model  
 63 ( $p < 2e-16$ , Table 3). There was strong evidence that Xgboost - Centre model outperformed all other models,  
 64 with Mixed effects RF + RE: centre performing worst, followed by Mixed Effects Xgboost + RE: centre  
 65 (CEM difference: -0.2605 and -0.2583, respectively). RF - Centre and RF + Centre (hot-encoded) had  
 66 similar performance (CEM difference: -0.0073) and have performance rankings immediately below Xgboost  
 67 – Centre, but above GPBoost.

58

59 Variable selected Xgboost



## 70 Adapted Rain plots

71 Model performance differences showed a similar overall pattern for variable selected compared to non-  
72 variable selected models (Figure S21). GPBoost showed lower AUC compared to other models, with the  
73 contrast in difference being greater for variable selected than for non-variable selected models. There was  
74 lower variation in F1 score for variable selected models, with all models outperforming GPBoost. Although  
75 mixed-effects models RF and Xgboost obtained a higher net benefit for variable selected models, the net  
76 benefit performance was markedly lower compared to other models than for non-variable selected models.  
77 A detailed report of individual metric results comprising the CEM is given in (Table S15).

## 78 Statistical Analyses

79 No extreme outliers were found. The CEM scores were normally distributed for all models, as assessed by  
80 Shapiro-Wilk's test ( $p > 0.05$ ). There was strong evidence of a difference across models  $p < 0.0001$ , except  
81 between: a) RF + Centre (hot-encoded) and RF - Centre; b) Mixed-effects RF and Mixed Effects Xgboost; c)  
82 Xgboost + Centre (hot-encoded) and Xgboost - Centre (Table S13 and Figure S20). Dunnett's test showed  
83 that there is minute but insignificant increase in performance between Xgboost + Centre (hot-encoded) and  
84 Xgboost - Centre model (CEM Difference: 0.0002,  $p = 0.94$ , Table S14). Xgboost - Centre significantly  
85 outperformed all other models except Xgboost + Centre (hot-encoded). The next best performing model was  
86 RF + Centre (hot-encoded) followed by RF - Centre, GPBoost and Mixed-effects RF. Mixed-effects  
87 Xgboost demonstrated the worst performance (CEM Difference: -0.2539).

38

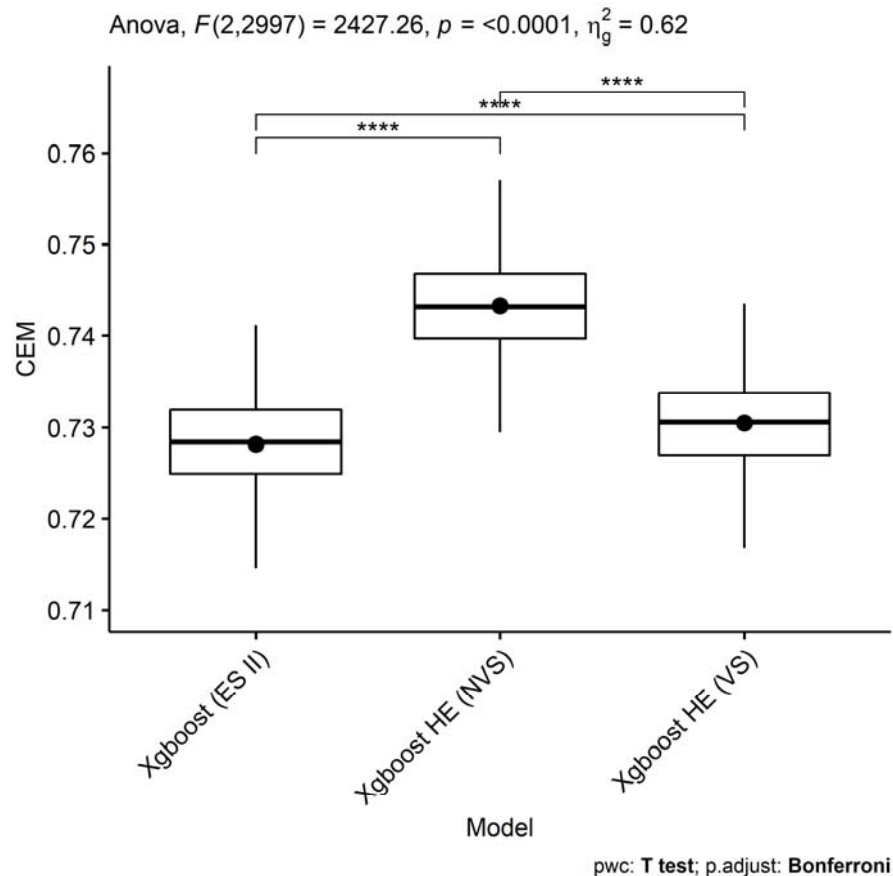
39 Highest performing: non-variable vs. variable selected model vs. Xgboost (ES II  
30 variables)

31 Xgboost + Centre (hot-encoded; 102 variables), Xgboost + Centre (hot-encoded; 18 variables) and Xgboost  
32 (ES II) variables models were compared. No extreme outliers were found. The CEM scores was normally  
33 distributed for all three models except Xgboost (ES II), as assessed by Shapiro-Wilk's test ( $p > 0.05$ ). A

34 histogram plot of the Xgboost (ES II) CEM values did not show substantial deviation from the normal  
35 distribution. There was strong evidence of a difference across models  $p < 0.0001$  (Table S16 and Figure 2b).

36

37 Figure 2. b) CEM performances of Xgboost (ES II), the best non-variable and variable selected models are  
38 compared using multiple pairwise paired t-tests with Bonferroni correction; NVS: non-variable selected; VS:  
39 variable selected; ES II: EuroSCORE II.



30

31

32 Dunnett's test showed that there was strong evidence that both non-variables selected (NVS) and variables  
33 selected (VS) Xgboost + Centre (hot-encoded) models were superior to the Xgboost (ES II) model ( $p < 6.3e-$   
34  $15$ , Table S17), with NVS Xgboost + Centre (hot-encoded) model having the best performance, followed by  
35 the VS Xgboost + Centre (hot-encoded) model (CEM Difference: 0.0150 and 0.0023, respectively).

## 36 Bayesian Networks

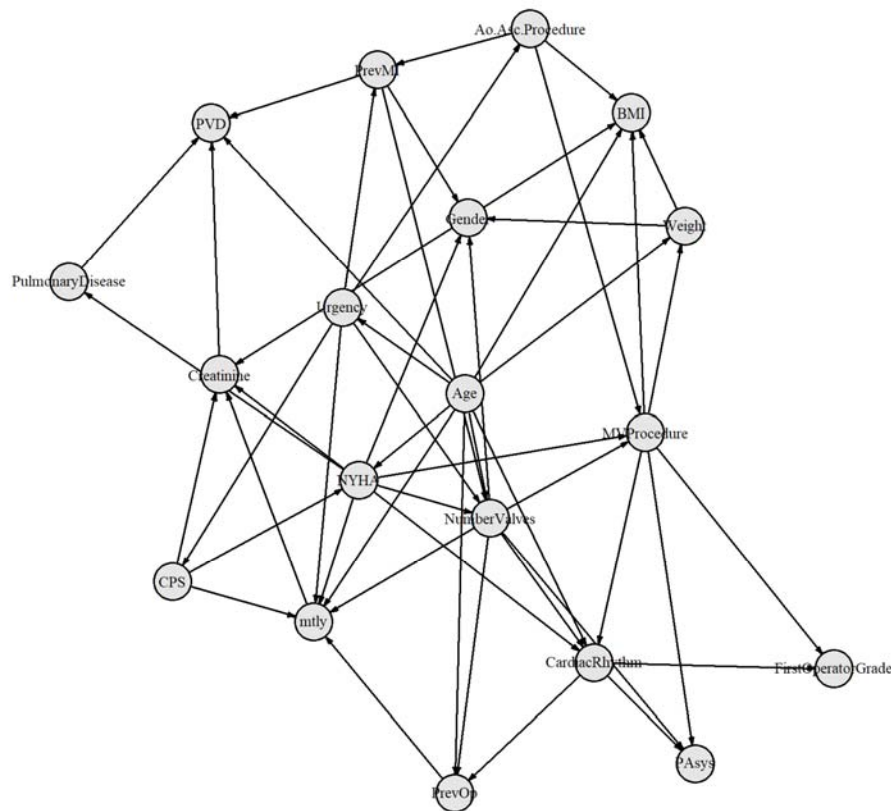
37 The Bayesian network for the interactions between the top 18 important variables, from the optimal model:

38 Xgboost + Centre (hot-encoded), and the outcome (mtly) shows that Urgency, Age, Critical Preoperative

39 State (CPS), NHYA, Number of Valves, Number of previous operations (PrevOp) have direct relationship /

10 path to the outcome variable (Figure 3). CPS confounds the relationship between NYHA and the outcome;  
11 Age confounds the relationship between Urgency and the outcome; NYHA confounds the relationship  
12 between Number of Valves and the outcome; Urgency confounds the relationship between CPS and the  
13 outcome; Number of Valves and Age both confounds the relationship between PrevOp and the outcome.  
14 Creatinine is a potential collider confounder that could bias the relationships between Urgency and the  
15 outcome.[39]

16 Figure 3. Bayesian network of interactions between the top 18 important variables, from Xgboost + Centre (hot-  
17 encoded) model, and the outcome, i.e. mtly: 0 – survival, 1 – non survival.



18  
19 The Bayesian network for the interactions between all the variables, from the optimal model:  
20 Xgboost + Centre (hot-encoded), and the outcome (mtly) shows that only Urgency and Cardiogenic Shock  
21 have a direct relationship/path to the outcome (Figure 4). Urgency confounds the relationship between  
22 Cardiogenic Shock and the outcome. Urgency mediates the relationship between Active Endocarditis  
23 (Endocarditis) and the outcome, whilst Endocarditis mediates the relationship between Mitral Valve



39 potentially be taken into account. However, these potential goldmines in clinical data usage have yet to be  
40 fully harvested. One example of this problem is exemplified in the controversial decision over whether to  
41 use a single procedural or a multi-procedural (universal) risk score/model.

42 While some studies have investigated the potential to devise scoring systems for specific surgical  
43 procedures such as tricuspid valve surgery using additional scores for other diseases such as a score for end-  
44 stage liver disease (MELD),[40] they have been limited by samples size of cohorts, unavailability of multi-  
45 centre data and limited use of holdout data. Conversely, universal (or general) risk scores have been  
46 developed focusing on a wider range of procedures than that by ES II and STS-PROM,[41] such as the  
47 American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) for  
48 mortality and morbidity for more than 100 different procedures.[42] Universal models such as the  
49 EuroSCORE and ES II have the benefit of allowing a large number of covariates to be included to improve  
50 model performance for low event rate (e.g. <1%) cardiac surgery datasets,[41] as well as allowing surgeons  
51 to evaluate the risk for nearly any combination of cardiac procedures, something that the procedure specific  
52 STS-PROM score does not allow.[41] There is mixed evidence as to whether a universal or procedure-  
53 specific score is preferable.[43]

54 Unlike our study, the current state of procedural specific and universal scores described above did  
55 not adjusted for hospital-based variation using a machine learning-based approach. This study also provides  
56 evidence for the use of larger number of covariates in universal cardiac surgery risk prediction models, and  
57 substantially increases the number of risk factors compared to only 18 variables of ES II. This was achieved  
58 by comparing ML models built from a combination of a large set of 61 variables, and adjusting for  
59 geographical variation as a result of 42 different hospital contributions. In addition, by using a larger multi-  
60 centre national dataset, the number of events (6,258; mortality rate of 2.76%) increased compared to smaller  
61 sized studies (<1%).[41] Therefore, despite using a substantially larger number of variables in the best  
62 performing Xgboost non-variable selected model with 102 variables, the Events Per Variable (EPV) rate  
63 remained ( $EPV = 6,258 / 102 = 61$ ) substantially larger than that previously found through simulation to be  
64 necessary for low bias and error in prediction modelling ( $EPV > 25$ ).[13]

35 This study provides strong evidence that the ES II variables are suboptimal for risk stratification  
36 modelling and that a larger set of 60 variables with additional adjustment for hospital variation provides  
37 superior performance to ML models built using ES II variables. The Xgboost model with hospital location  
38 adjustment of the large variable set performed best overall. Whilst the parsimonious models using the small  
39 (18) number of variables provided significantly weaker performance than the full variable set model, we  
40 demonstrated that there was significant evidence that the new small (18) variable models (Xgboost  
41 variations) outperformed the ES II 18 variable model (Xgboost). However, the improvements through  
42 adjustment for hospital location in the small variable set is so small through the dummy coding approach  
43 that this may not be necessary.

44 Notably, this study provides strong evidence that adjustment for hospital location improves ML  
45 model performance, but only in the presence of adjustment using a large set of variables and that cardiac  
46 surgery risk models with relatively small number of variables may not require this adjustment process. One  
47 possible reason for this is that in the small variable set models, more confounders are not being adjusted for.  
48 It is also worth noting that this effect applied to certain ML models, namely Xgboost, but did not apply to  
49 Random Forest. This work also shows that adjustment, whereby hospital location variables are separated  
50 into individual binary or dummy variables, demonstrated strong evidence of being superior over adjusting  
51 for location using a single vector random variable in a mixed-effects model. It is possible that separating out  
52 the hospital random effects into a multi-dimension set of individual vectors enabled the non-linear machine  
53 learning models to better adjust for the complex interactions across differences (or heterogeneity) in patient  
54 and operative characteristics across different cardiac centres.

55 One limitation of universal scores is that unadjusted variables that are not part of the score may  
56 negatively impact procedures that are more reliant on such variables, than other procedures, leading to  
57 inconsistency in procedural performances.[44] It has been highlighted that the high importance of certain  
58 interventional variables in risk scores is confounded by other factors, especially for the sickest patients,[11]  
59 making scores such as ES II and STS less useful for procedures for which confounding variables are not  
60 included.

31 This study is one of the first in its kind to demonstrate the importance of using Bayesian Networks  
32 (BN) for confounding analysis in the clinical setting as an alternative to the typically used Propensity score,  
33 which although useful, may overlook subtle causal relationships that could confound the real relationship  
34 between clinical variables and the outcome. The proposed BN approach enables clinicians to interpret the  
35 causal relationships in light of the confounding and is well suited to studies involving large variables such as  
36 the current study. One interesting observation is that by including the full set of variables in contrast to only  
37 the top 18 most important variables, many of the previously identified confounding relationships across  
38 variables disappeared. We identified that Cardiogenic Shock is the main cause of mortality following  
39 cardiac surgery, even though causation does not necessarily indicate optimal importance for prediction.  
40 Although Urgency (or highly urgent) procedures are also more likely to result in a higher risk of mortality, it  
41 could be seen that MV procedures for Endocarditis are likely to result in higher Urgency cases and  
42 consequently lead to a higher risk of surgical mortality. Another interesting finding is that the longer the  
43 interval between myocardial infarction (MI) and surgery, the higher the Urgency and consequently the  
44 higher the mortality risk. Clinical efforts should therefore target minimising the interval between MI and  
45 surgery. We also identified ascending aorta and specifically root procedure as high risk.

46 This study also provides a novel application of the adapted Rain plot for taking a wholistic view of  
47 the individual constituent metrics within the Clinical Effective Metric (CEM) consensus metric, and enables  
48 the filtering of high performance models. This provides a useful tool for clinicians to better understand how  
49 the CEM arrives at its ranking of competitive risk scores.

10

## 11 Limitations and Future work

12 This study is not without limitations. As the STS-PROM model coefficients are not made publicly available,  
13 we were not able to compare our universal model against a procedural specific model using this combination  
14 of new variables. Future studies should aim to compare the effects of pre-selecting features[18] using linear  
15 mixed effects adjustment of hospital variations in relation to the hot-encoding adjustment of random effects  
16 as well as considering multiple random effects including adjustment for surgeon differences.[17] To validate



17 the findings of this study, future studies may also compare the models considered herewithin using the linear  
18 mixed effects adjustment as a post-processing step.[15,16] Although the optimal models proposed in this  
19 study outperformed all models from our previous studies,[21,22] one possible reason for the mixed-effects  
20 models underperforming compared to other models is that mixed-effects ML models are typically better  
21 suited to regression tasks and are less well suited to classification tasks. Although the current study is mainly  
22 focused on the generalisability in terms of performance ranking on the hold out dataset, future work should  
23 investigate the effects of overfitting by considering approaches such as using, for example, the normalised  
24 ratio of holdout to training performance metric values. While a large number of clinically relevant variables  
25 have been taken into account, there may still be variables and hence residual confounding not adjusted for  
26 due to the availability of variables in the dataset. The effects of including or excluding variables such as  
27 Creatinine that demonstrated potential collider confounding relationships need to be further assessed.  
28 Further work is required to enhance the performance of mixed-effects models for the purpose of cardiac  
29 surgery risk classification.

30

## 31 CONCLUSION

32 This study based on a multi-centre national registry dataset comprising 42 UK hospitals highlights the  
33 importance of a larger set of potential confounding variables previously not considered by the EuroSCORE  
34 II. Furthermore, it suggests to adjust for hospital variation with specific recommendations for applying the  
35 ML model to study big data whereby each hospital is separated into individual binary input variables. We  
36 identified an ML-based hospital variation adjusted risk score comprising a large number of clinically  
37 predictive variables that increases risk stratification performance on hold out dataset, removing the need to  
38 perform variable selection and reduction. We demonstrated the concept of Bayesian Networks for cardiac  
39 surgery mortality associated causal relationship analysis following comparative risk score selection and  
40 identified Cardiogenic Shock, Urgency, the interval between myocardial infarction (MI) and surgery, MV  
41 procedures for Endocarditis, ascending aorta procedures and aortic root procedures to either directly or  
42 indirectly cause a higher risk of cardiac surgery mortality. Lastly, this study highlights the versatility of the



13 adapted Rain plot for rapid clinical assessment of which risk score should be considered for cardiac surgery  
14 patient selection across multiple consistent metrics and risk scores. It is recommended that this approach  
15 should be used in conjunction with the more robust CEM consensus metric and the Bayesian Network  
16 whereby possible. Future work will examine this new scoring approach in the context of performance drift  
17 and take into account procedural adjustments.

18

## 19 Funding

20 This work was supported by a grant from the BHF-Turing Institute and the NIHR Biomedical Research Centre at  
21 University Hospitals Bristol and Weston NHS Foundation Trust and the University of Bristol.

22

## 23 Contributorship

24

25 T.D., S.S., D.P.F., J.C., B.Z., P.N., M.C., U.B., A.J., A.D., G.D.A. contributed to experimental design. T.D. and S.S. acquired  
26 data. T.D. and S.S. performed the data preprocessing. T.D. wrote the source code to perform the experiments, and  
27 are accountable for all aspects of the work. T.D., S.S., D.P.F., J.C., B.Z., P.N., M.C., A.J., A.D., G.D.A. analyzed the results.  
28 T.D. wrote the first version of the paper. All authors revised the paper and approve the submission.

29

## 30 Data availability

31 All data used in this study are from the National Adult Cardiac Surgery Audit (NACSA) dataset. These data may be  
32 requested from Healthcare Quality Improvement Partnership (HQIP), [https://www.hqip.org.uk/national-](https://www.hqip.org.uk/national-programmes/accessing-ncapop-data/#.Ys6gN-zMLdp)  
33 [programmes/accessing-ncapop-data/#.Ys6gN-zMLdp](https://www.hqip.org.uk/national-programmes/accessing-ncapop-data/#.Ys6gN-zMLdp). Code for deriving training, update, and hold-out datasets is  
34 available on GitHub and authors can provide confirmatory de-identified record IDs for each set upon reasonable  
35 request.

36

## 57 Competing Interests

58 All authors declare that there are no competing interests.

59

## 70 Ethics statement

71 The study was approved by the Health Research Authority (HRA) and Health and Care Research Wales (HCRW) in 23  
72 of July 2019, IRAS project ID: 278171 and a waiver for patients' consent was obtained.

73 Guarantor TD.

## 74 Code availability

75 All source code used in this study are available on GitHub (<https://github.com/s0810110/EnsembleScoreAdaption>).

76 Analyses were performed using Scikit-learn v0.23.1, Keras v2.4.0, STATA-MP version 17 and R v4.0.2.

77

## 78 References

- 79 1 Ong CS, Reinertsen E, Sun H, *et al.* Prediction of operative mortality for patients undergoing cardiac surgical  
30 procedures without established risk scores. *The Journal of Thoracic and Cardiovascular Surgery* Published Online  
31 First: 14 September 2021. doi:10.1016/j.jtcvs.2021.09.010
- 32 2 Benedetto U, Dimagli A, Sinha S, *et al.* Machine learning improves mortality risk prediction after cardiac surgery:  
33 Systematic review and meta-analysis. *The Journal of Thoracic and Cardiovascular Surgery* Published Online First:  
34 10 August 2020. doi:10.1016/j.jtcvs.2020.07.105
- 35 3 Kieser TM, Rose MS, Head SJ. Comparison of logistic EuroSCORE and EuroSCORE II in predicting operative  
36 mortality of 1125 total arterial operations †. *European Journal of Cardio-Thoracic Surgery* 2016;**50**:509–18.  
37 doi:10.1093/ejcts/ezw072
- 38 4 Poullis M, Pullan M, Chalmers J, *et al.* The validity of the original EuroSCORE and EuroSCORE II in patients over the  
39 age of seventy. *Interactive CardioVascular and Thoracic Surgery* 2015;**20**:172–7. doi:10.1093/icvts/ivv345
- 30 5 Zhang G, Wang C, Wang L, *et al.* Validation of EuroSCORE II in Chinese Patients Undergoing Heart Valve Surgery.  
31 *Heart, Lung and Circulation* 2013;**22**:606–11. doi:10.1016/j.hlc.2012.12.012
- 32 6 Silaschi M, Conradi L, Seiffert M, *et al.* Predicting Risk in Transcatheter Aortic Valve Implantation: Comparative  
33 Analysis of EuroSCORE II and Established Risk Stratification Tools. *Thorac Cardiovasc Surg* 2015;**63**:472–8.  
34 doi:10.1055/s-0034-1389107
- 35 7 Carnero-Alcázar M, Silva Guisasola JA, Reguillo Lacruz FJ, *et al.* Validation of EuroSCORE II on a single-centre 3800  
36 patient cohort. *Interactive CardioVascular and Thoracic Surgery* 2013;**16**:293–300. doi:10.1093/icvts/ivs480

- 37 8 Arangalage D, Cimadevilla C, Alkhoder S, *et al.* Agreement between the new EuroSCORE II, the Logistic EuroSCORE  
38 and the Society of Thoracic Surgeons score: Implications for transcatheter aortic valve implantation. *Archives of*  
39 *Cardiovascular Diseases* 2014;**107**:353–60. doi:10.1016/j.acvd.2014.05.002
- 40 9 Atashi A, Amini S, Tashnizi MA, *et al.* External Validation of European System for Cardiac Operative Risk Evaluation  
41 II (EuroSCORE II) for Risk Prioritization in an Iranian Population. *Braz J Cardiovasc Surg* 2018;**33**:40–6.  
42 doi:10.21470/1678-9741-2017-0030
- 43 10 Provenchère S, Chevalier A, Ghodbane W, *et al.* Is the EuroSCORE II reliable to estimate operative mortality  
44 among octogenarians? *PLOS ONE* 2017;**12**:e0187056. doi:10.1371/journal.pone.0187056
- 45 11 Buratto E, Schulz A, Konstantinov IE. Commentary: Pulmonary complications in cardiothoracic surgery: Risk  
46 stratification is good; risk modification is better. *The Journal of Thoracic and Cardiovascular Surgery* 2021;**0**.  
47 doi:10.1016/j.jtcvs.2021.09.006
- 48 12 Peduzzi P, Concato J, Kemper E, *et al.* A simulation study of the number of events per variable in logistic  
49 regression analysis. *Journal of Clinical Epidemiology* 1996;**49**:1373–9. doi:10.1016/S0895-4356(96)00236-3
- 50 13 Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply  
51 related to events per variable. *J Clin Epidemiol* 2016;**76**:175–82. doi:10.1016/j.jclinepi.2016.02.031
- 52 14 Di Castelnuovo A, Bonaccio M, Costanzo S, *et al.* Common cardiovascular risk factors and in-hospital  
53 mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the  
54 multicentre Italian CORIST Study. *Nutrition, Metabolism and Cardiovascular Diseases* 2020;**30**:1899–913.  
55 doi:10.1016/j.numecd.2020.07.031
- 56 15 Bhuva AN, Bai W, Lau C, *et al.* A Multicenter, Scan-Rescan, Human and Machine Learning CMR Study to Test  
57 Generalizability and Precision in Imaging Biomarker Analysis. *Circulation: Cardiovascular Imaging*  
58 2019;**12**:e009214. doi:10.1161/CIRCIMAGING.119.009214
- 59 16 Gravesteijn BY, Nieboer D, Ercole A, *et al.* Machine learning algorithms performed no better than regression  
60 models for prognostication in traumatic brain injury. *Journal of Clinical Epidemiology* 2020;**122**:95–107.  
61 doi:10.1016/j.jclinepi.2020.03.005
- 62 17 Nasief H, Zheng C, Schott D, *et al.* A machine learning based delta-radiomics process for early prediction of  
63 treatment response of pancreatic cancer. *npj Precis Onc* 2019;**3**:1–10. doi:10.1038/s41698-019-0096-z
- 64 18 Parmley KA, Higgins RH, Ganapathysubramanian B, *et al.* Machine Learning Approach for Prescriptive Plant  
65 Breeding. *Sci Rep* 2019;**9**:17132. doi:10.1038/s41598-019-53451-4
- 66 19 Benedetto U, Sinha S, Lyon M, *et al.* Can machine learning improve mortality prediction following cardiac  
67 surgery? *European Journal of Cardio-Thoracic Surgery* 2020;**58**:1130–6. doi:10.1093/ejcts/ezaa229
- 68 20 S S, T D, A D, *et al.* Comparison of Machine Learning Techniques in Prediction of Mortality following Cardiac  
69 Surgery: Analysis of over 220,000 patients from a Large National Database. *European journal of cardio-thoracic*  
70 *surgery*: official journal of the European Association for Cardio-thoracic Surgery Published Online First: 5 August  
71 2023. doi:10.1093/ejcts/ezad183
- 72 21 Dong T, Sinha S, Zhai B, *et al.* Performance drift is a major barrier to the safe use of machine learning in  
73 cardiac surgery. 2023;:2023.01.21.23284795. doi:10.1101/2023.01.21.23284795
- 74 22 Angelini G, Dong T, Sinha S, *et al.* An Ensemble approach for Ensemble-Modelled Cardiac Surgery Risk  
75 Evaluation, Data Usage and Clinical Interpretability. 2022. doi:10.21203/rs.3.rs-1905463/v1
- 76 23 Hickey GL, Grant SW, Cosgriff R, *et al.* Clinical registries: governance, management, analysis and applications.  
77 *European Journal of Cardio-Thoracic Surgery* 2013;**44**:605–14. doi:10.1093/ejcts/ezt018

- 38 24 Dong T, Benedetto U, Sinha S, *et al.* Deep recurrent reinforced learning model to compare the efficacy of  
39 targeted local versus national measures on the spread of COVID-19 in the UK. *BMJ Open* 2022;**12**:e048279.  
40 doi:10.1136/bmjopen-2020-048279
- 41 25 StataCorp. Stata Statistical Software: Release 17. College Station, TX: StataCorp LLC; 2021.
- 42 26 Rajliwall NS, Davey R, Chetty G. Cardiovascular Risk Prediction Based on XGBoost. In: *2018 5th Asia-Pacific*  
43 *World Congress on Computer Science and Engineering (APWC on CSE)*. 2018. 246–52.  
44 doi:10.1109/APWCConCSE.2018.00047
- 45 27 Sarica A, Cerasa A, Quattrone A. Random Forest Algorithm for the Classification of Neuroimaging Data in  
46 Alzheimer’s Disease: A Systematic Review. *Front Aging Neurosci* 2017;**9**:329. doi:10.3389/fnagi.2017.00329
- 47 28 Barda N, Riesel D, Akriv A, *et al.* Developing a COVID-19 mortality risk prediction model when individual-level  
48 data are not available. *Nat Commun* 2020;**11**:4439. doi:10.1038/s41467-020-18297-9
- 49 29 Hickey GL, Blackstone EH. External model validation of binary clinical risk prediction models in cardiovascular  
50 and thoracic surgery. *The Journal of Thoracic and Cardiovascular Surgery* 2016;**152**:351–5.  
51 doi:10.1016/j.jtcvs.2016.04.023
- 52 30 Huang C, Li S-X, Caraballo C, *et al.* Performance Metrics for the Comparative Analysis of Clinical Risk  
53 Prediction Models Employing Machine Learning. [Miscellaneous Article]. *Circulation: Cardiovascular Quality &*  
54 *Outcomes* 2021;**14**. doi:10.1161/CIRCOUTCOMES.120.007526
- 55 31 Kumar NK, Sindhu GS, Prashanthi DK, *et al.* Analysis and Prediction of Cardio Vascular Disease using Machine  
56 Learning Classifiers. In: *2020 6th International Conference on Advanced Computing and Communication Systems*  
57 *(ICACCS)*. 2020. 15–21. doi:10.1109/ICACCS48705.2020.9074183
- 58 32 Tiwari P, Colborn KL, Smith DE, *et al.* Assessment of a Machine Learning Model Applied to Harmonized  
59 Electronic Health Record Data for the Prediction of Incident Atrial Fibrillation. *JAMA Network Open*  
50 2020;**3**:e1919396–e1919396. doi:10.1001/jamanetworkopen.2019.19396
- 51 33 Mehrtash A, Wells WM, Tempany CM, *et al.* Confidence Calibration and Predictive Uncertainty Estimation  
52 for Deep Medical Image Segmentation. *IEEE Transactions on Medical Imaging* 2020;**39**:3868–78.  
53 doi:10.1109/TMI.2020.3006437
- 54 34 Steyerberg EW, Vickers AJ, Cook NR, *et al.* Assessing the Performance of Prediction Models: A Framework for  
55 Traditional and Novel Measures. *Epidemiology* 2010;**21**:128–38. doi:10.1097/EDE.0b013e3181c30fb2
- 56 35 Allyn J, Allou N, Augustin P, *et al.* A Comparison of a Machine Learning Model with EuroSCORE II in Predicting  
57 Mortality after Elective Cardiac Surgery: A Decision Curve Analysis. *PLOS ONE* 2017;**12**:e0169772.  
58 doi:10.1371/journal.pone.0169772
- 59 36 Henglin M, Niiranen T, Watrous JD, *et al.* A Single Visualization Technique for Displaying Multiple  
70 Metabolite–Phenotype Associations. *Metabolites* 2019;**9**:128. doi:10.3390/metabo9070128
- 71 37 González-Estrada E, Cosmes W. Shapiro–Wilk test for skew normal distributions based on data  
72 transformations. *Journal of Statistical Computation and Simulation* 2019;**89**:3258–72.  
73 doi:10.1080/00949655.2019.1658763
- 74 38 Aragam B, Zhou Q. Concave Penalized Estimation of Sparse Gaussian Bayesian Networks. *Journal of Machine*  
75 *Learning Research* 2015;**16**:2273–328.
- 76 39 Tattan-Birch H, Marsden J, West R, *et al.* Assessing and addressing collider bias in addiction research: the  
77 curious case of smoking and COVID-19. *Addiction* 2021;**116**:982–4. doi:10.1111/add.15348

- 78 40 Färber G, Marx J, Scherag A, *et al.* Risk Stratification for Isolated Tricuspid Valve Surgery Assisted by Model of  
79 End-Stage Liver Disease Score. *The Journal of Thoracic and Cardiovascular Surgery* Published Online First: 12  
30 March 2022. doi:10.1016/j.jtcvs.2021.11.102
- 31 41 Tam DY, Fremes SE. Commentary: One size doesn't always fit all. *The Journal of Thoracic and Cardiovascular*  
32 *Surgery* 2020;**160**:180–1. doi:10.1016/j.jtcvs.2019.12.093
- 33 42 Bilimoria KY, Liu Y, Paruch JL, *et al.* Development and Evaluation of the Universal ACS NSQIP Surgical Risk  
34 Calculator: A Decision Aide and Informed Consent Tool for Patients and Surgeons. *J Am Coll Surg* 2013;**217**:833-  
35 842.e3. doi:10.1016/j.jamcollsurg.2013.07.385
- 36 43 Neumann F-J, Sousa-Uva M, Ahlsson A, *et al.* 2018 ESC/EACTS Guidelines on myocardial revascularization.  
37 *European Heart Journal* 2019;**40**:87–165. doi:10.1093/eurheartj/ehy394
- 38 44 Mori M, Shahian DM, Huang C, *et al.* Surgeons: Buyer beware—does “universal” risk prediction model apply  
39 to patients universally? *The Journal of Thoracic and Cardiovascular Surgery* 2020;**160**:176-179.e2.  
30 doi:10.1016/j.jtcvs.2019.11.144

31

32

33

34

35

36

37

38

39

30

31

32

### 33 **Figure legends**

34 Figure 1. Design overview of the study; Non-variable and variable selected analyses are performed; CEM is used to  
35 simultaneously assess discrimination, calibration, clinical utility, overall accuracy; analyses are performed for i)  
36 Mixed effects modelling with 60 main variables and hospitals as a single random variable; ii) Fixed effects modelling  
37 with 60 main variables and 42 hospitals each as a single variable and iii) Without consideration of hospital variation.

38 Table 1. Patient Demographics. Summary of cleaned and pooled top 27 most important variables from all models.  
39 Variables are for the time period 2012 – 2019. Records with missing mortality status were excluded.

10 Table 2. Non-variable selected models: Geometric Mean of Individual metrics; CEM refs to Clinical Effective Metric;  
11 Standard deviation and 95% CI are shown for CEM; adjusted 1 - ECE and 1 - Brier score values are shown; net benefit  
12 is average absolute overall benefit across all thresholds.

13 Figure 2. a) Non variable selected models: adapted Rain plot of CEM constituent metrics by model; larger sized  
14 spheres represent higher metric performance and vice versa.

15 Figure 2. b) CEM performances of Xgboost (ES II), the best non-variable and variable selected models are compared  
16 using multiple pairwise paired t-tests with Bonferroni correction; NVS: non-variable selected; VS: variable selected;  
17 ES II: EuroSCORE II.

18 Table 3. Non-variable selected models: Dunnett's test with Xgboost - Centre as control; 95% family-wise confidence  
19 level are shown as well as mean difference in CEM and p-values; NC: no centre; HE: hot-encoded centre; ME: mixed  
20 effects.

21 Figure 3. Bayesian network of interactions between the top 18 important variables, from Xgboost + Centre (hot-  
22 encoded) model, and the outcome, i.e. mtly: 0 – survival, 1 – non survival.

23 Figure 4 Bayesian network of interactions between all variables, from the non-variable selected Xgboost + Centre  
24 (hot-encoded) model, and the outcome (red), i.e. mtly: 0 – survival, 1 – non survival. The top 18 important variables  
25 are shown in green: within main network or in yellow: outside of the main network.

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44 Table 2. Patient Demographics. Summary of cleaned and pooled top 27 most important variables from all  
 45 models. Variables are for the time period 2012 – 2019. Records with missing mortality status were  
 46 excluded.

Variable	Mortality Status		p-value <sup>2</sup>
	0, N = 220,829 <sup>1</sup>	1, N = 6,258 <sup>1</sup>	
<b>Age (years), mean (SD)</b>	67.53 (11.23)	70.77 (11.42)	<0.001
<b>Female gender, n (%)</b>	59,467 (27%)	2,328 (37%)	<0.001
<b>Body Mass Index, mean (SD)</b>	28.49 (5.27)	27.93 (5.49)	<0.001
<b>Weight, mean (SD)</b>	82.42 (16.75)	78.98 (17.79)	<0.001
<b>Number of previous operations, n (%)</b>			<0.001
0	212,318 (96%)	5,288 (84%)	
1	7,600 (3.4%)	829 (13%)	
2	775 (0.4%)	105 (1.7%)	
3	115 (<0.1%)	26 (0.4%)	
4	19 (<0.1%)	9 (0.1%)	
5	1 (<0.1%)	0 (0%)	
6	1 (<0.1%)	1 (<0.1%)	
<b>Urgency, n (%)</b>			<0.001
0 - Elective	141,617 (64%)	2,442 (39%)	
1 - Urgent	72,090 (33%)	2,134 (34%)	
2 - Emergency	6,533 (3.0%)	1,230 (20%)	
3 - Salvage	589 (0.3%)	452 (7.2%)	
<b>First Operator Grade, n (%)</b>			<0.001
0 - Consultant	179,959 (81%)	5,729 (92%)	
1 - Associate specialist	6,726 (3.0%)	75 (1.2%)	
2 - Registrar/ SpR	30,113 (14%)	331 (5.3%)	
3 - SHO	4,031 (1.8%)	123 (2.0%)	

Variable	Mortality Status		
	0, N = 220,829 <sup>†</sup>	1, N = 6,258 <sup>†</sup>	p-value <sup>2</sup>
<b>Critical Preoperative State, n (%)</b>	7,255 (3.3%)	1,382 (22%)	<0.001
<b>Chronic pulmonary disease, n (%)</b>	26,644 (12%)	1,211 (19%)	<0.001
<b>Extra-cardiac Arteriopathy, n (%)</b>	22,327 (10%)	1,215 (19%)	<0.001
<b>Number of Grafts, n (%)</b>			<0.001
0	78,842 (36%)	2,935 (47%)	
1	18,320 (8.3%)	763 (12%)	
2	34,429 (16%)	880 (14%)	
3	60,648 (27%)	1,207 (19%)	
4	24,965 (11%)	397 (6.3%)	
5	3,294 (1.5%)	68 (1.1%)	
6	331 (0.1%)	8 (0.1%)	
<b>NYHA, n (%)</b>			<0.001
0 – I	48,625 (22%)	1,055 (17%)	
1 – II	96,888 (44%)	1,609 (26%)	
2 – III	64,049 (29%)	2,228 (36%)	
3 – IV	11,267 (5.1%)	1,366 (22%)	
<b>Creatinine, mean (SD)</b>	92.61 (47.15)	119.15 (84.48)	<0.001
<b>Cardiac Rhythm, n (%)</b>			<0.001
0 – Sinus	193,158 (87%)	4,677 (75%)	
1 – Preop AF	24,240 (11%)	1,309 (21%)	
2 – Preop VFT	399 (0.2%)	60 (1.0%)	
3 – Preop CHB or pacing	3,032 (1.4%)	212 (3.4%)	
<b>Number of Valves, n (%)</b>			<0.001
0	133,133 (60%)	2,941 (47%)	
1	77,621 (35%)	2,545 (41%)	
2	9,148 (4.1%)	659 (11%)	
3	916 (0.4%)	111 (1.8%)	
4	11 (<0.1%)	2 (<0.1%)	
<b>Pulmonary Artery Systolic Pressure, mean (SD)</b>	23.65 (11.87)	27.88 (16.80)	<0.001
<b>MV Procedure, n (%)</b>			<0.001
0 - None	191,803 (87%)	4,870 (78%)	
1 - Repair	17,633 (8.0%)	467 (7.5%)	



Variable	Mortality Status		
	0, N = 220,829 <sup>1</sup>	1, N = 6,258 <sup>1</sup>	p-value <sup>2</sup>
2 - Replacement	11,393 (5.2%)	921 (15%)	
<b>Number of previous Myocardial infarction, n (%)</b>			<0.001
0	150,612 (68%)	3,916 (63%)	
1	60,266 (27%)	1,887 (30%)	
2	9,951 (4.5%)	455 (7.3%)	
<b>Ascending aorta procedure, n (%)</b>	8,011 (3.6%)	769 (12%)	<0.001
<b>Previous valve surgery, n (%)</b>	5,701 (2.6%)	669 (11%)	<0.001
<b>Cardiogenic Shock, n (%)</b>	2,379 (1.1%)	840 (13%)	<0.001
<b>Pre-op Dialysis, n (%)</b>			<0.001
0 - None	217,789 (99%)	5,764 (92%)	
1 - No dialysis but AKI	853 (0.4%)	164 (2.6%)	
2 - AKI within 6wks of surgery needing dialysis	696 (0.3%)	157 (2.5%)	
3 - CKD dialysis	1,491 (0.7%)	173 (2.8%)	
<b>Days between LHC and cardiac surgery date, mean (SD)</b>	91.07 (858.72)	95.97 (468.56)	<0.001
<b>LVEF, n (%)</b>	54.01 (7.38)	51.43 (9.78)	<0.001
<b>Inotropes, n (%)</b>	2,114 (1.0%)	726 (12%)	<0.001
<b>Aortic arch procedure, n (%)</b>	1,286 (0.6%)	216 (3.5%)	<0.001
<b>Aortic root procedure, n (%)</b>	4,069 (1.8%)	387 (6.2%)	<0.001

<sup>1</sup>Mean (SD) or Frequency (%)

<sup>2</sup>Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test

AF - Atrial Fibrillation; VFT - Ventricular fibrillation or ventricular tachycardia; CHB - Complete heart block / paced; CKD - Chronic kidney disease; LHC - Left Heart Catheterization; LVEF - Left ventricular ejection fraction.

47

48

49

50

51

52 Table 2. Non-variable selected models: Geometric Mean of Individual metrics; CEM refs to Clinical Effective  
 53 Metric; Standard deviation and 95% CI are shown for CEM; adjusted 1 - ECE and 1 - Brier score values are shown;  
 54 net benefit is average absolute overall benefit across all thresholds.

Model Category	ECE	AUC	Brier	F1	Net Benefit	CEM Mean	CEM S.D	CEM CI Lower limit	CEM CI Upper limit
GPBoost ME	0.991	0.808	0.976	0.267	0.899	0.716	0.005	0.715	0.716
RF HE	0.995	0.849	0.977	0.284	0.906	0.734	0.005	0.733	0.734
RF ME	0.520	0.843	0.744	0.270	0.290	0.480	0.004	0.480	0.481
RF NC	0.996	0.848	0.977	0.284	0.906	0.734	0.005	0.733	0.734
Xgboost HE	0.997	0.855	0.977	0.300	0.908	0.743	0.005	0.743	0.744
Xgboost ME	0.520	0.835	0.744	0.278	0.291	0.483	0.003	0.482	0.483
Xgboost NC	0.997	0.854	0.977	0.296	0.908	0.741	0.005	0.741	0.741

55

56

57

58 Table 3. Non-variable selected models: Dunnett's test with Xgboost - Centre as control; 95% family-wise  
 59 confidence level are shown as well as mean difference in CEM and p-values; NC: no centre; HE: hot-encoded  
 60 centre; ME: mixed effects.

Group 1	Group 2	CEM Difference (1-2)	P Value	95% CI	
				Lower Bound	Upper Bound
GPBoost ME	Xgboost NC (Control)	-0.0252	<2e-16 ***	-0.0257	-0.0247
RF HE		-0.0073	<2e-16 ***	-0.0078	-0.0067
RF ME		-0.2605	<2e-16 ***	-0.2610	-0.2600
RF NC		-0.0073	<2e-16 ***	-0.0078	-0.0068
Xgboost HE		0.0024	<2e-16 ***	0.0019	0.0029
Xgboost ME		-0.2583	<2e-16 ***	-0.2589	-0.2578

51 Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

52

53

54

55

56