Machine Learning-Based Risk Profile Classification: A Case Study for Heart Valve Surgery

Ulrich Bodenhofer
LIT AI Lab & Institute of Bioinformatics
Johannes Kepler University Linz, Austria
bodenhofer@bioinf.jku.at

Bettina Haslinger-Eisterer
Dept. for Anesthesiology and Critical Care
Kepler University Clinic, Linz, Austria
Bettina.Haslinger-Eisterer@kepleruniklinikum.at

Alexander Minichmayer
Dept. for Anesthesiology and Critical Care
Kepler University Clinic, Linz, Austria
Alexander.Minichmayer@kepleruniklinikum.at

Georg Hermanutz
Institute of Bioinformatics
Johannes Kepler University Linz, Austria
hermanutz@bioinf.jku.at

Jens Meier
Dept. for Anesthesiology and Critical Care
Kepler University Clinic, Linz, Austria
Jens.Meier@kepleruniklinikum.at

Abstract

We employ machine learning to predict the 30-days mortality after heart valve surgeries from demographic and preoperative parameters. We achieve AUC values of almost 84%, while the standard EuroSCORE I provides an AUC of only slightly more than 70% for the given cohort. These results indicate (1) that state-of-the-art machine learning is superior to traditional risk models and (2) that calibrating models to specific institutions and surgical procedures allows for more accurate predictions that have the potential to improve medical decision making.

1 Introduction

Due to medical progress, even complex cardiac surgery is safer than ever before: typical mortality rates are in the low single-digit range, and nowadays the risk profile of most cardio-surgical procedures has to be judged as excellent. However, despite the existence of several risk scores, the specific risk of an individual patient is difficult to judge. For example, the most established scoring system for cardiac surgery, the European System for Cardiac Operative Risk Evaluation (EuroSCORE; [10, 11]), yields a relative risk for a patient population, but the individual risk of a patient remains elusive [8, 10, 13, 16]. Furthermore, despite its wide use, there are considerable limitations to the EuroSCORE, especially when used in subgroups that were not particularly considered during development of the EuroSCORE [14, 18, 19]. Detailed knowledge of the expected individual risk and anticipation of possible complications could allow for modification of perioperative procedures and, thereby, could improve patient safety and outcome significantly.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

One more additional drawback of most of the established risk models like EuroSCORE is the fact that they are based on risk factors that have been identified previously to play a major role for outcome prediction. However, this limitation probably underrates the potential importance of other factors and even more neglects the impact of mutual, possibly non-linear, interference of several factors in a given risk factor constellation. For example, EuroSCORE I and II are compiled from only 17–18 parameters that have been identified by a linear regression model.

Modern machine learning methods could overcome these shortcomings — by not relying on expert-selected features and linear dependencies like standardized scoring systems, such as EuroSCORE, do. Thereby, they have the potential to better describe the complex interaction between health risk factors in order to model the individual patient's risk — facilitating sensible models for more precise personalized risk prediction.

This paper describes a joint effort by machine learning researchers and clinicians: we use machine learning to predict individual risk of heart valve surgeries from preoperative clinical data. Thereby, we are able to devise a sensible cohort- and intervention-specific risk scoring system and to identify the main parameters that facilitate good predictions.

2 Material and Methods

This is a monocentric retrospective study (Kepler University Clinic, Linz, Austria; approved by ethical committee as study K-82-15) with one cohort consisting of data of patients who underwent heart valve surgery of any kind between January 1, 2008, and December 31, 2014. All data including survival data were obtained from the patients' anonymized electronic health records. The data set underwent extensive data pre-processing and data cleaning. The data cleaning included detection of missing values, typos, and out-of-range values. The original data set consisted of 2,229 patients and 147 features. After omission of intraoperative and postoperative features, omission of features with more than 25% missing data, imputation of the remaining missing values, and one-hot encoding of categorical features, the final data set contained 137 columns with preoperative input features.

We chose the 30-days mortality as the target value to be predicted, which is a commonly accepted standard in cardiac surgery [1, 6, 12]. In our study, the mortality rate within the first 30 days after surgery was 3.86% (86 of 2,229 cases).

In order to investigate the ability of machine learning to classify the 30-days mortality of patients undergoing valve surgery, nested five-fold cross validation was performed with hyperparameter selection on the four training folds. We employed the model selection procedure for three classes of the currently most popular and powerful machine learning methods: random forests (RFs), artificial neural networks (ANNs), and support vector machines (SVMs), where we considered three different prediction scenarios: (1) all features including EuroSCORE I; (2) all features except EuroSCORE I, in order to determine how well machine learning models are able to predict risk without including a risk score explicitly; (3) only preoperative parameters used by EuroSCORE I, in order to find out whether the parameters pre-selected for EuroSCORE I are sufficient for a decent prediction of mortality. For each of these scenarios, all three classes of machine learning methods have been considered. The models were constructed according to three distinct optimization criteria, accuracy (ACC), balanced accuracy (BACC), and area under the ROC curve (AUC). The two latter are necessary, since this is a highly unbalanced setting for which classification accuracy (ACC) is of limited value.

For random forest models [2, 3], hyperparameter selection was performed using out-of-bag estimates on the four training folds of the nested five-fold cross validation procedure. As hyperparameters, the number of features per split ($\lfloor \log_2(d) \rfloor$, $\lfloor \sqrt{d} \rfloor$, and $\lfloor d/3 \rfloor$ random features per split, where d denotes the number of features) and the sampling scheme (standard, balanced RF sampling scheme [4], and 3- and 10-fold oversampling of the smaller class) were considered.

When using artificial neural networks, hyperparameter selection was performed using four-fold cross validation on the respective training sets. We considered all possible combinations of network architectures (one hidden layer with 25, 50, 100, or 200 nodes and two hidden layers with 25, 50, and 100 nodes each), numbers of training epochs (150 and 300), learning rates (0.005 and 0.01), momentum (0.5 and 0.9), class weighting (overweighting of smaller class by a factor of 1, 7.5, and 15), input and hidden dropout (with probabilities 0.2 and 0.5), and L_2 weight decay (without or of

Table 1: Model selection results for all three prediction scenarios: the numbers show average performance measures on the five independent test folds (standard deviations shown in parentheses).

	Method	Crit.	AUC (%)	ACC (%)	BACC (%)	SENS (%)	SPEC (%)
with EuroSCORE I	RF	ACC	**83.00 (5.004)**	**96.64 (0.727)**	56.75 (4.148)	13.50 (8.297)	100.0 (0.000)
	RF	BACC	*79.88 (4.598)*	84.52 (1.151)	68.67 (5.551)	51.53 (12.04)	85.81 (1.340)
	RF	AUC	*82.96 (4.160)*	*96.50 (0.683)*	55.99 (3.525)	12.07 (7.123)	99.91 (0.209)
	ANN	ACC	74.20 (7.445)	*96.37 (0.750)*	58.55 (2.825)	17.57 (5.739)	99.53 (0.438)
	ANN	BACC	73.58 (8.671)	84.79 (5.223)	*71.23 (6.978)*	56.56 (12.95)	85.89 (5.304)
	ANN	AUC	78.01 (5.263)	*95.29 (0.794)*	64.18 (7.379)	30.51 (15.52)	97.85 (1.095)
	PSVM	ACC	*79.75 (5.846)*	*96.28 (0.341)*	51.50 (3.354)	3.000 (6.708)	100.0 (0.000)
	PSVM	BACC	*79.14 (4.215)*	71.42 (1.984)	**73.85 (1.596)**	76.49 (4.757)	71.20 (2.207)
without EuroSCORE I	RF	ACC	*81.91 (4.030)*	*96.59 (0.661)*	56.04 (3.561)	12.07 (7.123)	100.0 (0.000)
	RF	BACC	*79.44 (4.265)*	84.30 (1.619)	67.56 (2.813)	49.46 (7.031)	85.67 (1.749)
	RF	AUC	*82.05 (4.091)*	*96.55 (0.646)*	56.01 (3.539)	12.07 (7.123)	99.95 (0.104)
	ANN	ACC	71.52 (7.801)	*96.05 (0.605)*	57.40 (3.707)	15.50 (7.479)	99.30 (0.233)
	ANN	BACC	70.21 (11.35)	89.81 (4.893)	64.72 (6.059)	37.55 (14.41)	91.88 (5.415)
	ANN	AUC	77.70 (6.917)	94.57 (1.627)	65.64 (3.852)	34.25 (9.846)	97.02 (2.229)
	PSVM	ACC	77.51 (5.773)	*96.28 (0.341)*	51.50 (3.354)	3.000 (6.708)	100.0 (0.000)
	PSVM	BACC	*78.70 (4.629)*	71.20 (2.407)	*73.56 (4.759)*	76.14 (9.987)	70.98 (2.574)
EuroSCORE I features	RF	ACC	75.49 (10.30)	*96.10 (0.585)*	50.54 (1.343)	1.176 (2.631)	99.91 (0.128)
	RF	BACC	77.02 (10.74)	77.43 (2.749)	*70.87 (10.03)*	63.81 (21.17)	77.92 (3.050)
	RF	AUC	76.37 (11.50)	84.21 (9.916)	*68.28 (11.69)*	51.14 (34.31)	85.41 (11.42)
	ANN	ACC	73.19 (11.67)	*95.83 (0.628)*	50.40 (1.162)	1.176 (2.631)	99.63 (0.354)
	ANN	BACC	66.81 (11.54)	87.35 (1.859)	61.52 (4.564)	33.53 (10.28)	89.51 (2.341)
	ANN	AUC	70.37 (11.78)	*96.01 (0.582)*	49.93 (0.105)	0.000 (0.000)	99.86 (0.209)
	PSVM	ACC	52.29 (5.475)	*96.14 (0.486)*	50.00 (0.000)	0.000 (0.000)	100.0 (0.000)
	PSVM	BACC	68.89 (12.11)	80.44 (3.703)	62.25 (7.235)	42.57 (15.56)	81.94 (3.982)

with factor 0.001). For all ANNs we trained, we used ReLU activations [7] for the hidden nodes and a sigmoid for the output node.

All tests with regular support vector machines (SVMs; [5, 15]) were unsuccessful, because of the strong unbalancedness of the data set. Alternatively, we employed the Potential Support Vector Machine (PSVM; [9]) which offers a special balancing mode for optimizing balanced accuracy of classifiers trained on unbalanced data sets. PSVM hyperparameter selection was performed twice, once without balancing (to optimize for accuracy) and once with balancing (to optimize for balanced accuracy) using all combinations of cost factors $C \in \{8, 10, 12, \dots, 20\}$ and shrinkage parameters $\varepsilon \in \{0.5, 0.7, 0.9, \dots, 2.3\}$. PSVM was used in dyadic mode, i.e. the model's discriminant function is a linear combination of a subset of features.

3 Results

Table 1 summarizes the model selection results for all three prediction scenarios, all machine learning methods, and all model selection criteria. The values correspond to average performance measures on the five independent test sets, while standard deviations are shown in parentheses. The highest AUC values, accuracies, and balanced accuracies are shown in bold. Results that are insignificantly worse ($\alpha = 0.05$) than the best are shown in italics.

As obvious from Table 1, random forests give the best average classification accuracy on the data set with EuroSCORE I as a separate feature. A large number of other results from all three methods show accuracies that are only insignificantly worse than the maximal value. However, classification accuracy is not necessarily a very sensible measure in this unbalanced setting. The models optimized for balanced accuracy (rows with criterion BACC) obviously have much better sensitivities and worse specificities. The best average balanced accuracy is obtained for the data set with EuroSCORE I as a separate feature with PSVM using balancing. In such an unbalanced setting, in particular, when it is not immediately clear how the sensitivity-specificity trade-off should be addressed, AUC is a much more useful measure. Table 1 shows that the best average AUC is also obtained on the data set with EuroSCORE I as a separate feature. Whether or not EuroSCORE I is used, though, seems to have little influence: AUC values are slightly worse, but only insignificantly. Generally, AUCs

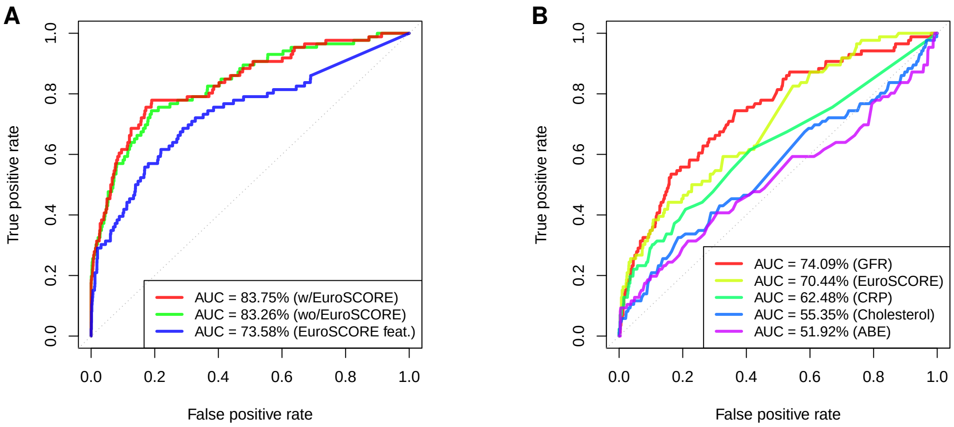

Figure 1: **A:** ROC curves computed from the out-of-bag predictions of random forests trained on our entire data set; **B:** ROC curves illustrating the predictiveness of the top five features.

appear to be best for random forest, followed by PSVM. Our ANNs, although providing competitive accuracies and balanced accuracies, cannot compete with random forest and PSVM in terms of AUC.

Generally, we see from Table 1 that results are often only insignificantly worse if EuroSCORE I is omitted. This is not surprising, since we use complex models that are able to implicitly "recreate" EuroSCORE I internally if necessary. However, it is obvious that restricting to the features used for EuroSCORE I leads to dramatically worse results. That clearly indicates that additional features that were not included in EuroSCORE I provide a clear advantage for risk assessment in our study.

We trained a random forest on our entire data set and all features including EuroSCORE I using the hyperparameter that have turned out to be the best strategy in the majority of cases in the cross validation procedure described above. This model gives an out-of-bag AUC of 83.75%. If we omit EuroSCORE I, we obtain an AUC of 83.26%, and if we restrict to EuroSCORE features, we obtain an AUC of 73.58% (see Fig. 1A). We also computed the importances of features in the random forests trained on the entire data set including and excluding EuroSCORE I. We observed that EuroSCORE I is the top feature if it is included. However, the omission of EuroSCORE I not only has little effect on the overall classification performance (see also above and Table 1), also the rankings of the other top features are the same, with arterial base excess (ABE), ICD-10 code I21 (i.e. acute myocardial infarction), cholesterol, C-reactive protein (CRP), and glomerular filtration rate (GFR) being the next most important features. We see from Fig. 1 that CRP, cholesterol, and ABE have relatively poor predictiveness by themselves. GFR, however, is more predictive than the accepted EuroSCORE. Note, however, that GFR is not among the features employed by EuroSCORE.

EuroSCORE and all other risk scores attribute continuous values to patients that can be understood as "the higher the score, the higher the risk". Our random forest classifiers yield probability estimates for the individual 30-days mortality which may serve as a proxy for the overall risk of patients analogous to EuroSCORE or other risk scores. The ROC curves in Fig. 1 underscore that the general ranking performance of these probability estimates are superior to EuroSCORE I. If we consider the 10% of patients with the highest risk as predicted by our random forest model, the overall 30-days mortality in the high-risk group is 22.42% (50 out of 223), while the 30-days mortality of the 10% of patients with the worst EuroSCORE I is 12.56% (28 out of 223). If we consider those 5% of patients with the highest risk as predicted by our random forest model, the overall 30-days mortality in this high-risk group is 29.46% (33 out of 112), while the 30-days mortality of the 10% of patients with the worst EuroSCORE I is 19.64% (22 out of 112). This underscores that our random forest predictor has a much higher precision when selecting high-risk patients than EuroSCORE I.

4 Discussion and Conclusion

Categorization of postoperative 30-day mortality with machine learning resulted in robust prediction models with an area under the ROC curve of almost 84%. Results were slightly superior when the calculated EuroSCORE I had been included into the model, whereas using solely the EuroSCOREassociated parameters resulted in a lower AUC. This is due to the inclusion of more relevant features (e.g. GFR) in our more general models. Note, however, that even if only the EuroSCORE features are used, non-linear models provide an advantage over the linear regression approach employed by EuroSCORE (compare Fig. 1A/EuroSCORE feat. with Fig. 1B/EuroSCORE).

In this study, the best results were obtained for a random forests. Neural networks were almost on a par. For larger data sets, deeper networks would become suitable which could then potentially outperform random forests by their ability to learn complex abstract representations.

Analyzing the most prominent parameters employed by the random forest classifiers, no simple cause becomes evident. All parameters, such as, CRP, cholesterol, ABE, and the most predictive GFR appear to be proxies for the patient's general condition with no or only loose causality.

A recent meta-analysis [17] has shown that the cumulated AUC of several studies for established risk models were generally lower than the AUCs given in the original publications. This is not a problem of the original studies per se, but simply caused by the fact that prediction models only deliver comparable results on a new cohort if this cohort has similar characteristics and fulfills the same assumptions as the cohort on which the prediction model was trained on. As our study demonstrates, if the data set is large enough, more precise risk models can be devised for the special conditions of a given institution and/or a given kind of surgery.

References

[1] B. Billah, C. M. Reid, G. C. Shardey, and J. A. Smith. A preoperative risk prediction model for 30-day mortality following cardiac surgery in an Australian cohort. *Eur. J. Cardiothorac. Surg.*, 37(5):1086–1092, 2010.

[2] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[3] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, editors. *Classification and Regression Trees*. CRC Press, 1984.

[4] C. Chen, A. Liaw, and L. Breiman. Using random forest to learn imbalanced data. Technical Report 666, Dept. of Statistics, University of California, Berkeley, 2004.

[5] C. Cortes and V. N. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1986.

[6] M.-B. Edwards and K. M. Taylor. Is 30-day mortality an adequate outcome statistic for patients considering heart valve replacement? *Ann. Thorac. Surg.*, 76:482–486, 2003.

[7] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proc. 14th Int. Conf. on Artificial Intelligence and Statistics*, pages 315–323, 2011.

[8] E. L. Hannan, H. Kilburn, M. Racz, E. Shields, and M. R. Chassin. Improving the outcomes of coronary artery bypass surgery in New York State. *JAMA–J. Am. Med. Assoc.*, 271(10):761–766, 1994.

[9] S. Hochreiter and K. Obermayer. Support vector machines for dyadic data. *Neural Comput.*, 18:1472–1510, 2006.

[10] S. A. M. Nashef, F. Roques, P. Michel, E. Gauducheau, S. Lemeshow, R. Salamon, and the EuroSCORE study group. European system for cardiac operative risk evaluation (EuroSCORE). *Eur. J. Cardiothorac. Surg.*, 16(1):9–13, 1999.

[11] S. A. M. Nashef, F. Roques, L. D. Sharples, J. Nilsson, C. Smith, A. R. Goldstone, and U. Lockowandt. EuroSCORE II. *Eur. J. Cardiothorac. Surg.*, 41(4):734–745, 2012.

[12] B. R. Osswald, E. H. Blackstone, U. Tochtermann, G. Thomas, C. F. Vahl, and S. Hagl. The meaning of early mortality after CABG. *Eur. J. Cardiothorac. Surg.*, 15(4):401–407, 1999.

[13] V. Parsonnet, D. Dean, and A. D. Bernstein. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. *Circulation*, 79(6):I3–I12, 1989.

[14] P. Pinna-Pintor, M. Bobbio, S. Colangelo, F. Veglia, M. Giammaria, D. Cuni, F. Maisano, and O. Alfieri. Inaccuracy of four coronary surgery risk-adjusted models to predict mortality in individual patients. *Eur. J. Cardiothorac. Surg.*, 21(2):199–204, 2002.

[15] B. Schölkopf and A. J. Smola. *Learning with Kernels*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2002.

[16] A. L. W. Shroyer, L. P. Coombs, E. D. Peterson, M. C. Eiken, E. R. DeLong, A. Chen, T. B. Ferguson, F. L. Grover, and F. H. Edwards. The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models. *Ann. Thorac. Surg.*, 75(6):1856, 2003.

[17] P. G. Sullivan, J. D. Wallach, and J. P. A. Ioannidis. Meta-analysis comparing established risk prediction models (EuroSCORE II, STS Score, and ACEF Score) for perioperative mortality during cardiac surgery. *Am. J. Cardiol.*, 118(10):1574–1582, 2016.

[18] M. van Gameren, A. P. Kappetein, E. W. Steyerberg, A. C. Venema, E. A. J. Berenschot, E. L. Hannan, A. J. J. C. Bogers, and J. J. M. Takkenberg. Do we need separate risk stratification models for hospital mortality after heart valve surgery? *Ann. Thorac. Surg.*, 85(33):921–930, 2008.

[19] D. Wendt, B. R. Osswald, K. Kayser, M. Thielmann, P. Tossios, P. Massoudy, M. Kamler, and H. Jakob. Society of Thoracic Surgeons score is superior to the EuroSCORE determining mortality in high risk patients undergoing isolated aortic valve replacement. *Ann. Thorac. Surg.*, 88(2):468–474, 2009.