Heliyon 9 (2023) e17139

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Temporal validation of 30-day mortality prediction models for transcatheter aortic valve implantation using statistical process control – An observational study in a national population

Ricardo R. Lopes [a,b,c], Tsvetan T.R. Yordanov [d,e], Anita A.C.J. Ravelli [d,e], Saskia Houterman [f], Marije Vis [g,h], Bas A.J.M. de Mol [c,i], Henk Marquering [a,b,h], Ameen Abu-Hanna [d,e,*], on behalf of the THI Registration Committee of the Netherlands Heart Registration

[a] Amsterdam UMC Location University of Amsterdam, Biomedical Engineering and Physics, Meibergdreef 9, Amsterdam, the Netherlands
[b] Amsterdam UMC Location University of Amsterdam, Radiology and Nuclear Medicine, Meibergdreef 9, Amsterdam, the Netherlands
[c] Amsterdam Cardiovascular Sciences, Heart Failure and Arrhythmias, Amsterdam, the Netherlands
[d] Amsterdam UMC Location University of Amsterdam, Medical Informatics, Meibergdreef 9, Amsterdam, the Netherlands
[e] Amsterdam Public Health Research Institute, Amsterdam, the Netherlands
[f] Netherlands Heart Registration, Utrecht, the Netherlands
[g] Amsterdam UMC Location University of Amsterdam, Cardiology, Meibergdreef 9, Amsterdam, the Netherlands
[h] Amsterdam Cardiovascular Sciences, Atherosclerosis and Ischemic Syndromes, Amsterdam, the Netherlands
[i] Amsterdam UMC Location University of Amsterdam, Cardiothoracic Surgery, Meibergdreef 9, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Transcatheter aortic valve implantation
Aortic stenosis
Prediction models
Machine learning
Temporal validation
Statistical process control

ABSTRACT

Background: Various mortality prediction models for Transcatheter Aortic Valve Implantation (TAVI) have been developed in the past years. The effect of time on the performance of such models, however, is unclear given the improvements in the procedure and changes in patient selection, potentially jeopardizing the usefulness of the prediction models in clinical practice. We aim to explore how time affects the performance and stability of different types of prediction models of 30-day mortality after TAVI.

Methods: We developed both parametric (Logistic Regression) and non-parametric (XGBoost) models to predict 30-day mortality after TAVI using data from the Netherlands Heart Registration. The models were trained with data from 2013 to the beginning of 2016 and pre-control charts from Statistical Process Control were used to analyse how time affects the models' performance on independent data from the mid of 2016 to the end of 2019. The area under the Receiver Operating Characteristics curve (AUC) was used to evaluate the models in terms of discrimination and the Brier Score (BS), which is related to calibration, in terms of accuracy of the predicted probabilities. To understand the extent to which refitting the models contribute to the models' stability, we also allowed the models to be updated over time.

Results: We included data from 11,291 consecutive TAVI patients from hospitals in the Netherlands. The parametric model without re-training had a median AUC of 0.64 (IQR 0.54–0.73) and BS of 0.028 (IQR 0.021–0.035). For the non-parametric model, the median AUC was 0.63 (IQR 0.48–0.68) and BS was 0.027 (IQR 0.021–0.036). Over time, the developed

* Corresponding author. Department of medical informatics, Amsterdam UMC - Location AMC, University of Amsterdam, P.O. Box 22660, 1105 AZ Amsterdam, the Netherlands.
E-mail address: a.abu-hanna@amsterdamumc.nl (A. Abu-Hanna).

https://doi.org/10.1016/j.heliyon.2023.e17139
Received 5 March 2023; Received in revised form 2 June 2023; Accepted 8 June 2023
Available online 10 June 2023

parametric model was stable in terms of AUC and unstable in terms of BS. The non-parametric model was considered unstable in both AUC and BS. Repeated model refitting resulted in stable models in terms of AUC and decreased the variability of BS, although BS was still unstable. The refitted parametric model had a median AUC of 0.66 (IQR 0.57–0.73) and BS of 0.027 (IQR 0.020–0.035) while the non-parametric model had a median AUC of 0.66 (IQR 0.57–0.74) and BS of 0.027 (IQR 0.023–0.035).

Conclusions: The temporal validation of the TAVI 30-day mortality prediction models showed that the models refitted over time are more stable and accurate when compared to the frozen models. This highlights the importance of repeatedly refitted models over time to improve or at least maintain their performance stability. The non-parametric approach did not show improvement over the parametric approach.

1. Introduction

Aortic stenosis is the most common valvular disease in developed countries. If symptomatic, the stenosis requires valve intervention [1]. Transcatheter Aortic Valve Implantation (TAVI) has become the routine treatment for aortic stenosis even for low and intermediate risk patients [2–4]. Besides the improvements of the procedure and technology involved [5,6], such as using smaller sheaths and organizing specialized teams for the procedure, a strict patient selection is being followed to select patients who are likely to benefit from TAVI [7].

The TAVI candidate selection is performed by a multi-disciplinary team, where multiple risk scores, such as STS (Society of Thoracic Surgery) [8] and EuroSCORE [9,10] are considered. Although those scores are not TAVI specific, they are well accepted parametric models and used for early-mortality estimation after cardiac surgery. Other instruments, such as FRANCE2 [11], ACC-TAVI [12], and also non-parametric models [13,14] aimed at predicting mortality specifically after TAVI have been introduced. The external validation of such models, in which patients originate from other settings and countries, has shown worse performance than on the internal validation obtained on the original dataset [15–17]. Such models only achieved improved performance when refitted (or in general updated) for that specific centre [18,19].

The TAVI patient selection process and the procedure itself are changing over time, and it is still unknown if there is a performance drift in the accuracy of the mortality prediction models over time in the same setting. With that, it is not clear if the prediction of models developed a while ago are stable and fit for continuous use without refitting. Although a limited prospective validation was performed in a previous study [18], only a single test set was used by the authors and the performance change and model's stability were not assessed repeatedly over time. In addition, the evaluated risk scores were developed using a parametric model and it is unclear how non-parametric models (such as boosting trees) behave on the same TAVI mortality prediction task. Therefore, an investigation is needed to assess the stability of models over long periods of time. Statistical Process Control (SPC) is a monitoring and alerting instrument that combines graphical and statistical inferences that can be used to monitor the accuracy and errors of the prediction models over time [20,21]. With this approach, the model's stability over time can be visualized and statistically assessed. We aim to explore how time affects the performance and stability of both a parametric and a non-parametric prediction model for 30-day mortality after TAVI. To this end, we used a large dataset from all heart centres in the Netherlands to train the models and use SPC to monitor their stability and performance prospectively.

2. Methods

2.1. Study population

We included all patients registered in the Netherlands Heart Registration (NHR)[1] who underwent a TAVI procedure between January 2013 and December 2019 in the Netherlands. The NHR is a national registry that includes data from all the sixteen-heart intervention centres in the Netherlands, containing demographics, clinical characteristics, intervention, and procedure details [22]. The NHR Transcatheter Heart Valve Interventions registration committee gave permission for this analysis in January 2021.

For this study, the outcome used is the 30-day mortality after the TAVI procedure. Two of the sixteen centres were excluded given that less than 5% of the 30-day mortality status of their patients was available when conducting the study. In addition, patients without a mortality status or that had a concomitant procedure (e.g., pacemaker implantation) were not included.

In order to analyse the studied population, general statistics were computed for all variables. Mean and standard deviation was computed for data with normal distribution and median and interquartile range for non-normally distributed data. Chi-square or Two-sample T-test was used as appropriate.

2.2. Variables

We included all variables that were available in the NHR and that had been already used in TAVI risk scores and other studies

[1] https://nederlandsehartregistratie.nl.

[13–19]. Among the variables, demographic data such as age, sex, and body mass index (BMI) were included. Also, clinical history and screening variables, including the estimated Glomerular Filtration Rate (eGFR), the New York Heart Association (NYHA) score, chronic lung disease, dialysis, systolic pulmonary arterial pressure, creatinine, diabetes mellitus (DM), left ventricular ejection fraction, and recent myocardial infarction were included. In terms of the procedure, its acuity, the chosen access route, critical preoperative state and, year of the procedure were included. All used variables were acquired before the TAVI procedure was performed.

The eGFR and creatinine were clipped for values larger than 60 mL/min/1.73 m^2 and 250 μmol/L, respectively based on expert opinion. Also, DM was represented by three categories: no DM, with untreated DM, and with DM being treated with insulin. Additionally, the procedure acuity and access route were dichotomized to elective/non-elective and femoral/non-femoral access respectively.

To deal with the missing values, an iterative multiple imputation method (MissForest) was used to impute data. For this step, only the data from the training set was used to train the imputation model, which was later used to impute the data on the test set. Dummy variables were created by leaving one category out for the NYHA score, year of the procedure, and DM categories.

2.3. Prediction models

We evaluated two well-established parametric and non-parametric techniques: logistic regression (LR) and extreme gradient boosting (XGB). LR is a parametric approach and has one coefficient assigned for each variable of the model, allowing a relatively easy interpretation and low model complexity. On the other hand, XGB is a non-parametric approach, based on building an ensemble of decision trees. With that, predictions of multiple trees are combined into a single prediction. The models were developed using the scikit-learn [23] and XGBoost [24] Python libraries.

Both models had their hyperparameters tuned using a grid-search approach with the training data in a stratified 10-fold cross-validation (CV). Specifically, different sets of parameters were assessed to find the optimum model, such as the error used for training the LR model, and the tree depth for XGB. All hyperparameters assessed are listed in the Supplementary Material Table S1. The hyperparameter set with the highest average Area Under the Curve (AUC) of the receiver operating characteristics curve across the tuning data, which is held out of the training set, of all folds was selected and used to train the models. In order to visualize the agreement between predicted mortality risk and real mortality, a calibration plot was created for all prediction models.

The reporting methods of this analysis adhere to the reporting guidelines (TRIPOD). The statement can be found in the Supplementary Material Table S2.

2.4. Model validation

2.4.1. Internal validation

Internal validation was performed to evaluate the models regardless of any temporal shifts in the data and have forms a reference for the temporal analysis. To this end, the data from all the treatment years were gathered together and a 10-fold CV imputation model was created based on the training folds and later used to impute the corresponding conducted as described above. The imputation model was created based on the training folds and later test set.

The average AUC, with standard deviation (SD), was used to evaluate the models. While the AUC is commonly used for the evaluation of clinical models, it is not sensitive to changes in the prevalence of the event. Hence, the Brier Score (BS), which is sensitive to prevalence and calibration was also selected as a measure of the accuracy of the predicted probabilities. The higher the AUC and the lower the BS, the better.

2.4.2. Temporal validation

Temporal validation was conducted to simulate the models' predictive performance over time, reproducing how they would perform if used in a real-life scenario with prospective patients. To this end, all patients were gathered together and sorted by their procedure date. We require a sufficient number of points to be monitored (about 25) and use the formula limit $n > 9*(1 - p)/p$, where p is the average (mortality percentage in our case). This formula is applied in the more rigorous p-chart to ensure a positive lower control. In our case $p = 3.36\%$ requiring a sample size of at least 239 patients. The data was split into 38 mutually exclusive groups. Except for the first group, with 320 samples, all remaining groups had 297 samples each.

2.5. Group analysis and pre-control charts

For context, we first visualize changes in 30-day mortality ratio and age over time. These are the most important and intuitive variables. Then, we prepare the data for the SPC analysis using pre-control charts. SPC is a graphical framework in which progression of a key measurement is plotted over time and, additionally, provides rules to judge the stability of the process by assessing whether the variation in the measurements reflects expected natural variation or a structural change. In our case, the analysed process represents the performance in terms of the AUC and Brier score of the parametric and non-parametric TAVI mortality prediction models over time. A structural change indicated instability. There are various types of control charts.

In this study, we use pre-control charts (also called zone charts) which show the process along with acceptable control limits on a graph. Their aim is to monitor a process for early detection of shifts, allowing for prompt corrective actions. Although they can be less reliable than the more elaborate traditional control charts, they are much more intuitive to use and can be effective, especially in low-risk (manufacturing) processes. Zone charts divide the chart into three zones based on specification limits in a "traffic light" design.

The green zone, defined by mean ± 2 SDs of the process, indicates a stable process. The yellow (warning) zone, within 2–4 SDs of the mean, indicates a stable process if no two or more consecutive points fall in this zone. The red zone, >4 SDs, indicates an unstable process if any point falls in this zone.

The acceptable limits can be pre-specified, like in industrial applications. The process can be run already after five consecutive points fall in the green zone to ensure that the process is stable and then normally about 25 points are monitored [25]. For a prediction model, the acceptable performance limits can be derived from the model's early demonstrated performance. Because we have enough data points, for 38 groups, and to include enough variability in the process in order to lower type 1 errors, we require 8 consecutive points (i.e., groups) to fall in the green zone in this study before we calculate the mean and standard deviation of the process, which are used in judging the nature of the variation in the remaining points. This evaluation strategy, starting from a stable process, is more challenging than fitting the model say on data of the first year and monitoring the process where it could be shown to be unstable already in its first monitored points.

Two experiments were performed: a) without re-training the model on each iteration frozen model and b) re-training the model on each successive iteration (model refitting). This is done to compare the stability between the fixed model and a model with repeated refitting over time. For the frozen model, the model was trained once and evaluated on all subsequent parts individually. The model refitting, on the other hand, has the testing data from the previous iteration added to the training data on each new iteration and has all the coefficients re-estimated. Fig. 1 shows a representation of both experiments. All statistical analysis were performed with Python (version 3.8.8). To implement the models, the scikit-learn (version 0.24.1) and XGBoost (version 1.3.3) libraries were used.

3. Results

In total, data from 12,440 TAVI patients matched our inclusion period 2013–2019 and were considered for this study. For the analysis, data from 11,291 patients were included after excluding 837 patients for not having 30-day mortality information, 309 for belonging to the two centres with a high missing rate of mortality information, and 3 patients for having an additional procedure (i.e., not isolated TAVI).

The mean age of the included patients was 79.72 ± 6.86 and 50.21% of the patients were female. The baseline and procedural characteristics of the population used in this study, as well as the descriptive statistics, can be found in Table 1. Mean mortality is 410/(410 + 10881) corresponding to 3.36%.

3.1. Internal validation

In the internal validation, with the inclusion of all 11,291 patients at the same time and a 10-fold CV, both the LR and XGB achieved a mean AUC of 0.68 and, respectively, a mean BS of 0.034 and 0.036 (Table 2). The calibration plots are available in the Supplementary Material Fig. S1.

In Fig. 2, the 30-day mortality and age of the patients are plotted over time. They demonstrate downward trends. When preparing data for the temporal validation for the pre-control charts (visible in the Supplementary Material Fig. S2), we observed that the first 4 performance points (2013–2014), also showed a trend, and were hence excluded. The subsequent 8 groups did show stable performance and hence were used to train the frozen model and the initial model that will subsequently be refitted. This left 26 points for monitoring.

Fig. 3 displays the performance of the LR and XGB frozen models. While the AUC was considerably stable for the LR model and remained stable after 2017 for the XGB, BS was mostly in the red zone (>4 SD) for both models. Fig. 4 displays the progress of performance over time when using the model refitting approach (note that the zone limits are continuously updated as well). Both LR and XGB models were stable in their AUC, but instability in BS is observed at the beginning. The AUC limits of the refitted LR model slightly changed compared to the frozen model. The AUC of the XGB model and BS of both models had their range visibly increased. This indicates a larger standard deviation which reflects larger uncertainty detected over time. The frozen parametric model had a median AUC of 0.64 (IQR 0.54–0.73) and BS of 0.028 (IQR 0.021–0.035) while the frozen non-parametric model had a median AUC of 0.63 (IQR 0.48–0.68) and BS of 0.027 (IQR 0.021–0.036). Regarding model refitting, the parametric model had a median AUC of 0.66 (IQR 0.57–0.73) and BS of 0.027 (IQR 0.020–0.035) while the non-parametric had a median AUC of 0.66 (IQR 0.57–0.74) and BS of 0.027

Fig. 1. Schematic representation of the experiments with a frozen model and model refitting scenario. The frozen model was kept unchanged for all iterations while model refitting implied re-training the model in every new iteration.

Table 1
Characteristics of the 11,291 TAVI patients stratified by their 30-day mortality survival status. Values are represented as mean and standard deviation (SD), median and interquartile range (IQR), number (n), or percentage (%).

		Grouped by 30-day mortality			
		Missing	Non-surv.	Surv.	p-value
n			410	10881	
Age (yr), mean (SD)		0	80.3 (7.2)	79.7 (6.9)	0.095
Sex, n (%)	Male	0	191 (46.6)	5430 (49.9)	0.205
	Female		219 (53.4)	5451 (50.1)	
BMI (kg/m2), mean (SD)		143	26.5 (5.6)	27.3 (4.9)	0.010
Year of procedure, n (%)	2013	0	60 (14.6)	723 (6.6)	<0.001
	2014		61 (14.9)	973 (8.9)	
	2015		55 (13.4)	1305 (12.0)	
	2016		57 (13.9)	1450 (13.3)	
	2017		60 (14.6)	1898 (17.4)	
	2018		60 (14.6)	2073 (19.1)	
	2019		57 (13.9)	2459 (22.6)	
eGFR (mL/min/1.73m2), mean (SD)		36	55.3 (21.6)	60.5 (29.3)	<0.001
NYHA class, n (%)	1	1151	32 (8.8)	1067 (10.9)	<0.001
	2		62 (17.1)	2718 (27.8)	
	3		216 (59.7)	5377 (55.0)	
	4		52 (14.4)	616 (6.3)	
Chronic lung disease, n (%)	No	37	298 (73.8)	8627 (79.5)	0.006
	Yes		106 (26.2)	2223 (20.5)	
Procedure acuity, n (%)	Elective	220	312 (80.6)	9733 (91.1)	<0.001
	Emergency		4 (1.0)	19 (0.2)	
	Urgent		71 (18.3)	932 (8.7)	
Dialysis, n (%)	No	212	387 (97.7)	10,570 (98.9)	0.042
	Yes		9 (2.3)	113 (1.1)	
TAVI access route, n (%)	Direct aortic access	15	56 (13.7)	781 (7.2)	<0.001
	Other access		2 (0.5)	15 (0.1)	
	Subclavian access		31 (7.6)	623 (5.7)	
	Transapical		53 (13.0)	679 (6.2)	
	Transf., other		37 (9.1)	684 (6.3)	
	Transf., percutaneous		179 (43.9)	6104 (56.2)	
	Trasnf., surgical		50 (12.3)	1982 (18.2)	
Critical preoperative state, n (%)	No	94	382 (96.2)	10,747 (99.5)	<0.001
	Yes		15 (3.8)	53 (0.5)	
Systolic pulmonary arterial pressure (mmHg), median (IQR)		2084	25.0 [25.0,38.5]	25.0 [25.0,31.2]	0.001

Non-surv: Non-survival, Surv: Survival, BMI: Body Mass Index, NYHA: New York Heart Association Functional Classification, TAVI: Transcatheter Aortic Valve Implantation, SD: Standard Deviation, IQR: Interquartile Range.

Table 2
Evaluation of the models trained without temporal assessment (internal validation) with standard deviation. *AUC* = Area Under the Receiver Operating Characteristic curve, *BS* = Brier Score.

Model/Metric	AUC	BS
Logistic Regression	0.68 ± 0.07	0.034 ± 0.001
XGBoost	0.68 ± 0.05	0.036 ± 0.001

Group analysis and pre-control charts.

(IQR 0.023–0.035).

Fig. 5 shows the calibration curves (all points combined), with the frozen and model refitting approaches, for both LR and XGB models. The frozen LR and XGB models are completely overestimating the predicted mortality risk. The model refitting approach does achieve a more balanced calibration. The calibration plots, assessed over time, are available in the Supplementary Material Fig. S3 (for LR) and Supplementary Material Fig. S4 for (XGB). Additionally, the selected hyperparameters are available in Supplementary Material Table S3 and Table S4.

4. Discussion

Without repeated refitting over time, the parametric TAVI mortality prediction model was considered stable regarding discrimination (in terms of the AUC) but unstable regarding the accuracy of the predicted probabilities (in terms of the BS). The non-parametric model was unstable in both AUC and BS. When models were repeatedly refitted over time, both parametric and non-parametric models

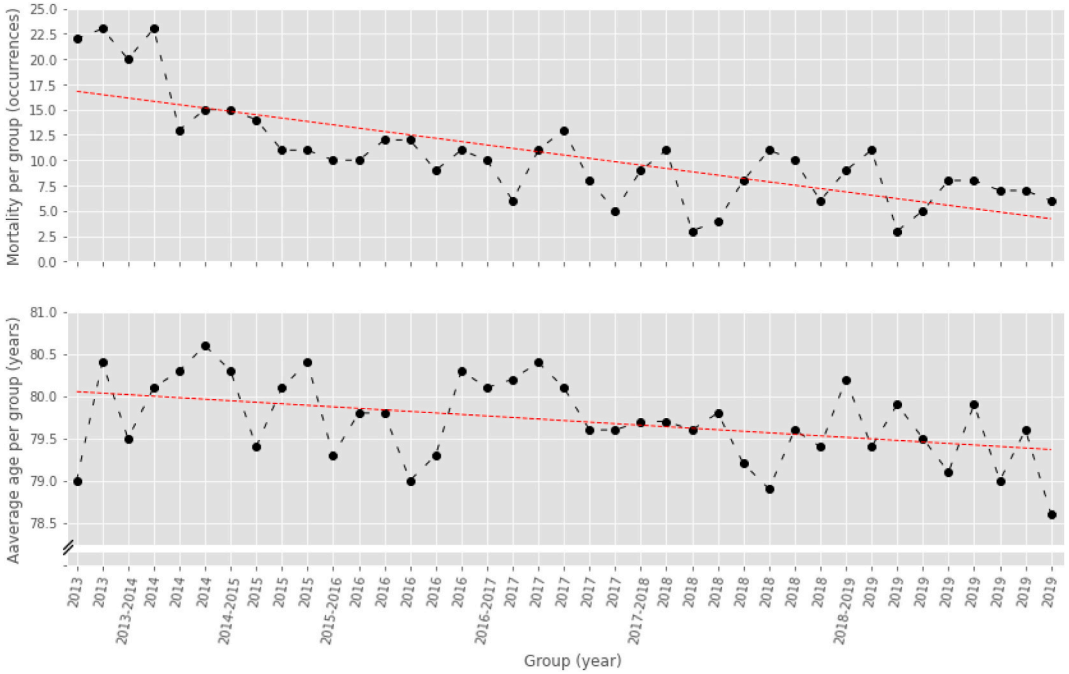

Fig. 2. Mean 30-day mortality and age over time of TAVI patients. A linear trend is presented in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

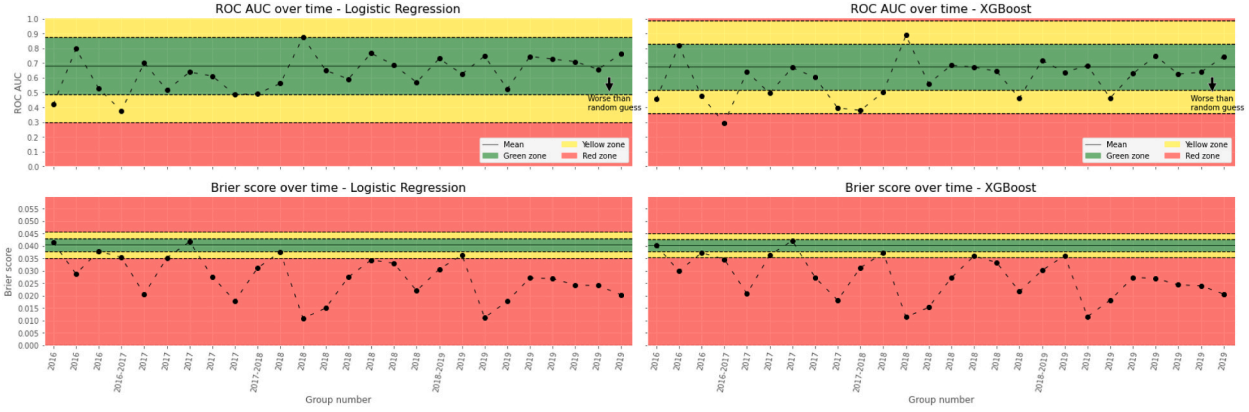

Fig. 3. Temporal validation of the frozen LR (left) and XGB (right) models. For the LR model, the AUC is considered stable and most of the BS points are inside the red zone, hence the BS is unstable. Regarding the XGB model, an AUC point and most of the BS points are inside the red zone, hence the process is unstable. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

were considerably more stable and had only a few points in the yellow and red zones.

TAVI procedures are over time offered to younger patients and patients with lower risk. Therefore, the mortality outcomes improved over time and the average age and mortality of the analysed TAVI population declined over time. When not refitted, this decline in mortality results in the tendency of the prediction models to overestimate the mortality probability, which in turn leads to unstable models. When we performed the model refitting, which also updates the limits, a widening in these limits was clearly visible reflecting the larger uncertainty because of the higher variation in the data. It is important to note that the AUC is much less affected by the mortality prevalence than the BS. This explains why the BS become quickly unstable in the frozen models.

Regarding TAVI mortality prediction models, Al-Farra et al. [18] performed a prospective analysis of mortality prediction models and highlighted the importance of performing model refitting to overcome performance drifts. In this latter study, two parametric models were analysed and the prospective data was treated as a single dataset, while we divided the prospective data into multiple groups and we used SPC. Also, recent studies using national registries from Germany and Switzerland [26,27] analysed temporal trends over the TAVI procedures, confirming the reduction in mortality we found. However, the accuracy and stability of the risk scores over time were not considered in these studies, nor was SPC used. Using pre-control charts to investigate stability over time on

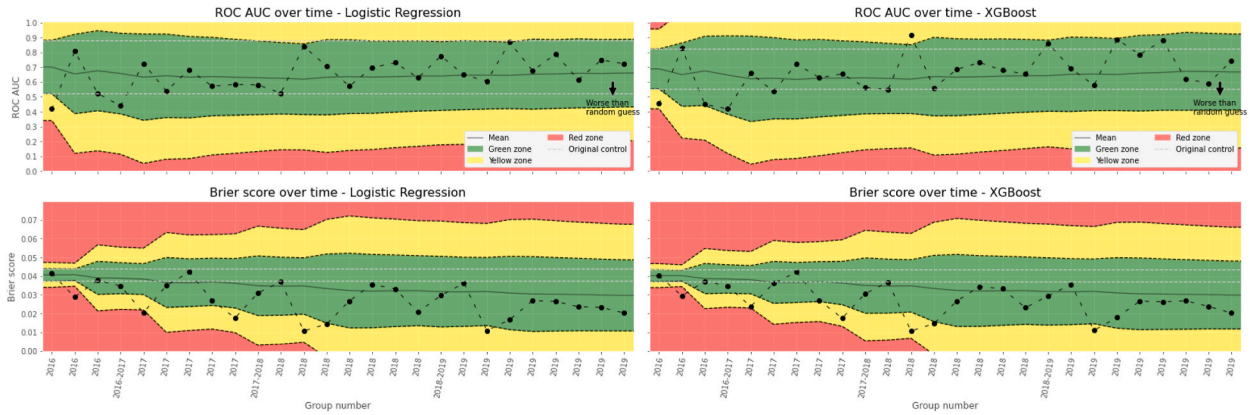

Fig. 4. Temporal validation of the model refitting for LR (left) and XGB (right). While the AUC is stable for both models, some of the BS points are in the red zone at the beginning, hence the BS is unstable. Note that the zone limits are recalculated per model refitting on a successive group. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

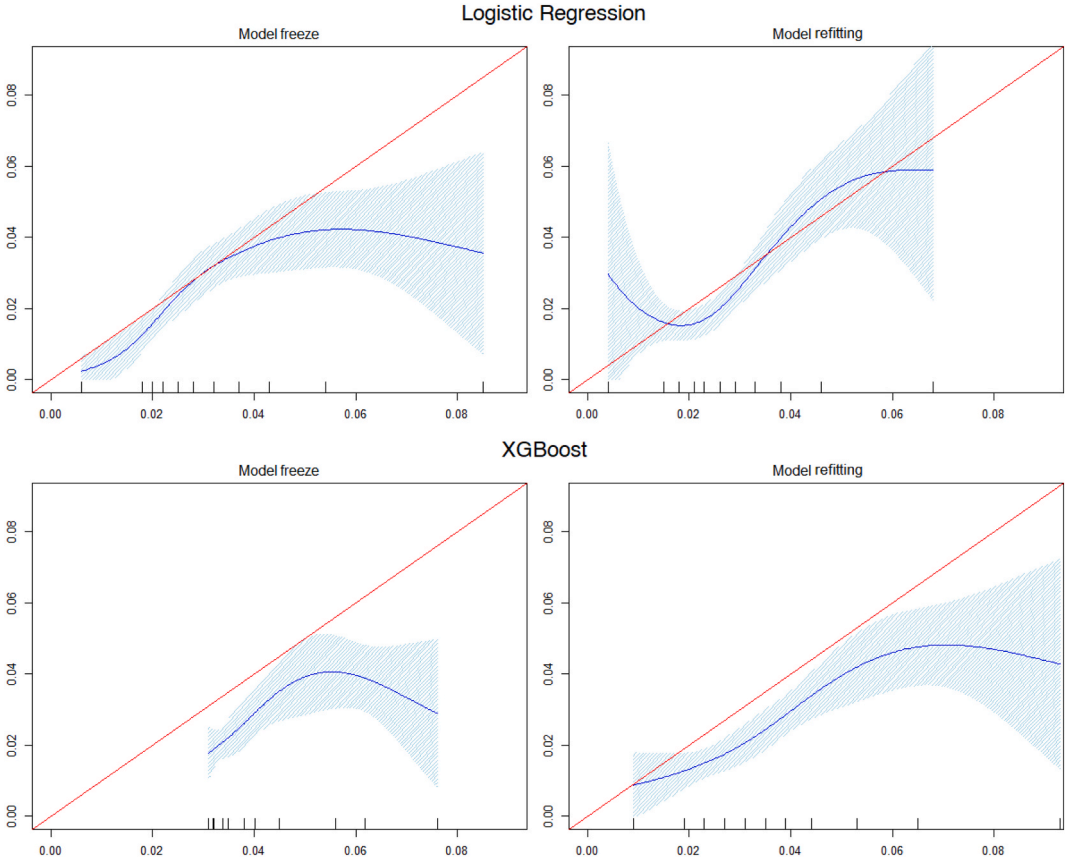

Fig. 5. Calibration plots of the LR and XGB models. The plots were generated after the combination of all data points.

prediction models had been used by Minne et al. [28,29] for evaluating pre-existent models for the prediction of mortality in the intensive care unit. Similar to our results, they found a significant difference within BS over time, while the AUC remained stable. However, they did not find time trends in the mortality or age of the observed patients. Also, the authors used a first-level recalibration approach, instead of re-training the model, to deal with the effects of time on the data. Although effective in their study, it was also suggested that more rigorous approaches, such as model refitting that we used, might be needed.

Strengths of this study include the use of a large national registry with more than 10,000 patients, with real recent data over many years. In addition, we compared two methods: a parametric (LR) and a non-parametric method (XGB). Furthermore, instead of simply

analysing a frozen model with prospective data, we proposed a model refitting approach and evaluated its performance. Finally, we used two important performance measures: AUC for gauging discrimination and the Brier score for measuring the accuracy of the predicted probability. We also looked at the implications of (in)stability in terms of calibration graphs. As far as we know, this is the first study performing temporal analysis of TAVI mortality prediction models with such techniques.

This study also has some limitations. This data is from a national registry and has multiple centres. The centres might have different standards for patient selection or the performance of the procedure and this information was not taken into account directly (the centre was not used as a feature). In addition, the analysed data is from a country with a mainly Caucasian population, so a country-specific analysis. Also, a fixed number of samples per group was used to better understand how the models change over time and this leads to a different number of groups per year/month. Considering the clinical implementation of this study, one would have to wait until the number of procedures is reached to include a new group. Regarding the technical implementation, there are several approaches that could have been used to update the coefficients dynamically [30,31] that were not explored in this initial study. Finally, in this work, we only validated the models on one future subgroup but did not attempt to evaluate it on the accumulated subgroups until a specified time.

Our work shows the importance of taking time into account when using mortality prediction models. Specifically, in our large dataset, the stability of both parametric and non-parametric models was considered poor, mainly for the BS. This demonstrates the danger of only considering AUC when evaluating prediction models, which is a common practice, and the importance of analysing multiple metrics when evaluating models. With model refitting, the stability increased for both parametric and non-parametric models. However, this improved stability came at the cost of more uncertainty in performance. We found that it might be risky to use a model for longer periods without refitting (or updating in general), independent of whether it is a parametric or non-parametric model. The frozen models were poorly calibrated and, also with model refitting, the calibration was still insufficient. An underestimation or overestimation of the predicted probability is seen in the calibration plots for both models. Finally, we would like to highlight the importance of inspecting the confidence intervals (reflecting honesty in uncertainty) rather than the absolute improvement of the models' performance.

Future work can investigate differences between centres. Also, in order to avoid the necessity of having enough patients to compose a new group, Individual Control Charts, which are able to analyse individual measurements, could be explored. In addition, a subgroup analysis could provide insight into specific groups that markedly diverge from the rest of the population. Furthermore, one hypothesis is that the older data might harm the model and can be ignored, or given less weight, over time once it might be too different from the current population. Moreover, a different evaluation strategy can be explored in which the frozen model is incrementally validated on the accumulated subgroups until some time instead of validating on one subgroup. Finally, reproducing this experiment with a different (TAVI) population or risk scores, such as STS or EuroSCORE, warrants further research.

5. Conclusion

In our study, the prediction models that were refitted over time were more stable and accurate compared to the frozen models. It highlights the importance of repeatedly refitting (or, in general, updating) the models over time to improve their performance stability. Although the refitted models were more stable, the calibration was still poor and it came also at the cost of more uncertainty in performance. There were no clear benefits in using the non-parametric model over the parametric model. The trained models, when not refitted, were unstable and presented a higher overestimation of 30-day mortality after TAVI than the models that were refitted over time.

Author contribution statement

Ricardo Ricci Lopes: Conceived and designed the experiments; Performed the experiments; Analysed and interpreted the data; Wrote the paper.

Tsvetan T.R. Yordanov: Conceived and designed the experiments; Analysed and interpreted the data.

Anita A.C.J. Ravelli, Saskia Houterman, Marije M. Vis, Bas A.J.M. de Mol, Henk Marquering: Analysed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Ameen Abu-Hanna: Conceived and designed the experiments; Analysed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

Mr Ricardo Ricci Lopes was supported by ITEA3 {16017}.

Data availability statement

The authors do not have permission to share data.

Addendum

The following physicians are the members of the NHR THI Registration Committee. They represent the hospitals that have provided

the data for this study. Contact with NHR THI Registration Committee can be via the e-mail info@nederlandsehartregistratie.nl

P. den Heijer - Amphia Hospital
M. V. Vis - Amsterdam University Medical Center, location AMC
W. A.L. Tonino - Catharina Hospital
N. M.D.A. van Mieghem - Erasmus University Medical Center
C. E. Schotborgh - Haga Hospital
V. Roolvink - Isala Hospital
F. van der Kley - Leiden University Medical Center
S. Kats - Maastricht University Medical Center
F. Porta - Medical Center Leeuwarden
M. G. Stoel - Medisch Spectrum Twente
G. Amoroso - OLVG Hospital
M. van Wely - Radboud University Medical Center
L. Timmers - St. Antonius Hospital
M. Voskuil - University Medical Center Utrecht
H. W. van der Werf - University Medical Center Groningen.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e17139.

References

[1] B.A. Carabello, W.J. Paulus, Aortic stenosis, Lancet 373 (9667) (2009) 956–966.
[2] H.H. Nielsen, Transcatheter aortic valve implantation, Dan. Med. J. 59 (12) (2012) B4556.
[3] F. Fang, J. Tang, Y. Zhao, J. He, P. Xu, A. Faramand, Transcatheter aortic valve implantation versus surgical aortic valve replacement in patients at low and intermediate risk: a risk specific meta-analysis of randomized controlled trials, PLoS One 14 (9) (2019), e0221922.
[4] C.A. Gomez, J. Braghiroli, E. de Marchena, "The Changing Paradigm": TAVR for Low-risk Patients Approved by the FDA, Wiley Online Library, 2020.
[5] M. Akodad, T. Lefèvre, TAVI: simplification is the ultimate sophistication, Front. Cardiovasc. Med. 5 (2018) 96.
[6] H.K. Abdelaziz, D.H. Roberts, Advances in transcatheter aortic valve implantation, in: W. Ahmed, D.A. Phoenix, M.J. Jackson (Eds.), Charalambous CPBT-A in M and SE, Academic Press, 2020, pp. 103–119.
[7] R.A. Nishimura, C.M. Otto, R.O. Bonow, B.A. Carabello, J.P. Erwin, R.A. Guyton, et al., 2014 AHA/ACC guideline for the management of patients with valvular heart disease: a report of the American college of cardiology/American heart association task force on practice guidelines, J. Am. Coll. Cardiol. 63 (22) (2014) e57–e185.
[8] S.M. O'Brien, D.M. Shahian, G. Filardo, V.A. Ferraris, C.K. Haan, J.B. Rich, et al., The society of thoracic surgeons 2008 cardiac Surgery risk models: Part 2-isolated valve Surgery, Ann. Thorac. Surg. 88 (1 SUPPL) (2009) S23–S42, https://doi.org/10.1016/j.athoracsur.2009.05.056.
[9] S.A.M. Nashef, F. Roques, P. Michel, E. Gauducheau, S. Lemeshow, R. Salamon, et al., European system for cardiac operative risk evaluation (Euro SCORE), Eur. J. Cardio-thoracic Surg. 16 (1) (1999) 9–13.
[10] S.A.M. Nashef, F. Roques, L.D. Sharples, J. Nilsson, C. Smith, A.R. Goldstone, et al., Euroscore ii, Eur. J. Cardio-thoracic Surg. 41 (4) (2012) 734–745.
[11] B. Iung, C. Laouénan, D. Himbert, H. Eltchaninoff, K. Chevreul, P. Donzeau-Gouge, et al., Predictive factors of early mortality after transcatheter aortic valve implantation: individual risk assessment using a simple score, Heart 100 (13) (2014) 1016–1023.
[12] F.H. Edwards, D.J. Cohen, S.M. O'Brien, E.D. Peterson, M.J. Mack, D.M. Shahian, et al., Development and validation of a risk prediction model for in-hospital mortality after transcatheter aortic valve replacement, JAMA Cardiol 1 (1) (2016) 46–52.
[13] R.R. Lopes, M.S. van Mourik, E.V. Schaft, L.A. Ramos, J. Baan, J. Vendrik, et al., Value of machine learning in predicting TAVI outcomes, Neth. Heart J. 27 (9) (2019) 443–450. Available from: http://link.springer.com/10.1007/s12471-019-1285-7.
[14] P. Agasthi, H. Ashraf, S.H. Pujari, M.E. Girardo, A. Tseng, F. Mookadam, et al., Artificial intelligence trumps TAVI2-SCORE and CoreValve score in predicting 1-year mortality post transcatheter aortic valve replacement, Cardiovasc. Revasc. Med. 24 (2020) 33–41.
[15] H. Al-Farra, A. Abu-Hanna, B.A.J.M. de Mol, W.J. Ter Burg, S. Houterman, J.P.S. Henriques, et al., External validation of existing prediction models of 30-day mortality after transcatheter aortic valve implantation (TAVI) in The Netherlands heart registration, Int. J. Cardiol. 317 (2020) 25–32.
[16] G. Wolff, J. Shamekhi, B. Al-Kassou, N. Tabata, C. Parco, K. Klein, et al., Risk modeling in transcatheter aortic valve replacement remains unsolved: an external validation study in 2946 German patients, Clin. Res. Cardiol. (2020) 1–9.
[17] G.P. Martin, M. Sperrin, P.F. Ludman, M.A. de Belder, C.P. Gale, W.D. Toff, et al., Inadequacy of existing clinical prediction models for predicting mortality after transcatheter aortic valve implantation, Am. Heart J. 184 (2017) 97–105.
[18] H. Al-Farra, B.A.J.M. de Mol, A.C.J. Ravelli, W. Ter Burg, S. Houterman, J.P.S. Henriques, et al., Update and, internal and temporal-validation of the France-2 and ACC-TAVI early-mortality prediction models for Transcatheter Aortic Valve Implantation (TAVI) using data from The Netherlands heart registration (NHR), IJC Hear. Vasc. 32 (2021), 100716.
[19] R.R. Lopes, M. Mamprin, J.M. Zelis, P.A.L. Tonino, M.S. van Mourik, M.M. Vis, et al., Inter-center cross-validation and finetuning without patient data sharing for predicting transcatheter aortic valve implantation outcome, in: 2020 IEEE 33rd Int Symp Comput Med Syst, 2020, pp. 591–596. Available from: https://ieeexplore.ieee.org/document/9183069/.
[20] L. Minne, S. Eslami, N. de Keizer, E. de Jonge, S.E. de Rooij, A. Abu-Hanna, Statistical process control for validating a classification tree model for predicting mortality–a novel approach towards temporal validation, J. Biomed. Inf. 45 (1) (2012) 37–44.

[21] L. Minne, S. Eslami, N. de Keizer, E. de Jonge, S.E. de Rooij, A. Abu-Hanna, Statistical process control for monitoring standardized mortality ratios of a classification tree model, Methods Inf. Med. 51 (4) (2012) 353–358.
[22] M.J.C. Timmermans, S. Houterman, E.D. Daeter, P.W. Danse, W.W. Li, Lipsic Erik, et al., Using real-world data to monitor and improve quality of care in coronary artery disease: results from The Netherlands Heart Registration [cited 2022 Apr 24]. Available from: Neth. Heart J. (2022) 1–9 https://link.springer.com/article/10.1007/s12471-022-01672-0.
[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2012) 2825–2830.
[24] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min., 2016, pp. 785–794.
[25] J.L. Smith, Pre-control may be the solution [acessed 2022 May 6]. Available from: https://www.qualitymag.com/articles/86794-pre-control-may-be-the-solution.
[26] V. Mauri, M. Abdel-Wahab, S. Bleiziffer, V. Veulemans, A. Sedaghat, M. Adam, et al., Temporal trends of TAVI treatment characteristics in high volume centers in Germany 2013–2020 [cited 2022 Mar 10]. Available from: Clin. Res. Cardiol. (2021) 1–8 https://link.springer.com/article/10.1007/s00392-021-01963-3.
[27] S. Stortecky, A. Franzone, D. Heg, D. Tueller, S. Noble, T. Pilgrim, et al., Temporal trends in adoption and outcomes of transcatheter aortic valve implantation: a SwissTAVI Registry analysis, Eur. Hear. J. – Qual. Care Clin. Outcomes 5 (3) (2019) 242–251 [cited 2022 Mar 10]. Available from: https://academic.oup.com/ehjqcco/article/5/3/242/5124351.
[28] L. Minne, S. Eslami, N. De Keizer, E. De Jonge, S.E. De Rooij, A. Abu-Hanna, Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment, Intensive Care Med. 38 (1) (2012) 40–46 [cited 2022 Mar 28]. Available from: https://pubmed.ncbi.nlm.nih.gov/22042520/.
[29] L. Minne, S. Eslami, N. de Keizer, E. de Jonge, S.E. de Rooij, A. Abu-Hanna, Statistical process control for monitoring standardized mortality ratios of a classification tree model, Methods Inf. Med. 51 (4) (2012) 353–358 [cited 2022 Mar 28]. Available from: https://pubmed.ncbi.nlm.nih.gov/22773038/.
[30] D.A. Jenkins, M. Sperrin, G.P. Martin, N. Peek, Dynamic models to predict health outcomes: current status and methodological challenges, Diagnostic Progn. Res. 2 (1) (2018) 1–9, 2018 Dec 18 [cited 2023 Mar 5]. Available from: https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-018-0045-2.
[31] T.L. Su, T. Jaki, G.L. Hickey, I. Buchan, M. Sperrin, A review of statistical updating methods for clinical prediction models, Stat. Methods Med. Res. 27 (1) (2016) 185–197 [cited 2023 Mar 5]. Available from: https://journals.sagepub.com/doi/10.1177/0962280215626466.