International Journal of Cardiology 317 (2020) 25–32

Contents lists available at ScienceDirect

International Journal of Cardiology

journal homepage: www.elsevier.com/locate/ijcard

External validation of existing prediction models of 30-day mortality after Transcatheter Aortic Valve Implantation (TAVI) in the Netherlands Heart Registration

Hatem Al-Farra [a,b,*,1], Ameen Abu-Hanna [a,1], Bas A.J.M. de Mol [b,1], W.J. ter Burg [a,1], Saskia Houterman [c,1], José P.S. Henriques [b,1], Anita C.J. Ravelli [a,1],on behalf of the NHR THI Registration Committee:

M.M. Vis [a], J. Vos [b], J. Ten Berg [c], W.A.L. Tonino [d], C.E. Schotborgh [e], V. Roolvink [f], F. Porta [g], M. Stoel [h], S. Kats [i], G. Amoroso [j], H.W. van der Werf [k], P.R. Stella [l], P. de Jaegere [m]

[a] Amsterdam University Medical Centers
[b] Amphia Hospital
[c] St. Antonius Hospital
[d] St. Antonius Hospital
[e] HagaHospital
[f] Isala
[g] Leeuwarden Medical Center
[h] Medisch Spectrum Twente
[i] Maastricht University Medical Center
[j] Onze Lieve Vrouwe Gasthuis
[k] University Medical Center Groningen
[l] University Medical Center Utrecht
[m] Erasmus University Medical Center

[a] Department of Medical Informatics, Amsterdam UMC - Location AMC, University of Amsterdam, the Netherlands
[b] Heart Center, Amsterdam UMC - Location AMC, University of Amsterdam, Amsterdam Cardiovascular Sciences, Amsterdam, the Netherlands
[c] The Netherlands Heart Registration, Utrecht, the Netherlands

ARTICLE INFO

Article history:
Received 19 December 2019
Received in revised form 19 April 2020
Accepted 13 May 2020
Available online 22 May 2020

Keywords:
Transcatheter aortic valve implantation (TAVI)
Mortality
Prediction model
External validation
Discrimination
Calibration

ABSTRACT

Background: Several mortality prediction models (MPM) are used for predicting early (30-day) mortality following transcatheter aortic valve implantation (TAVI). Little is known about their predictive performance in external TAVI populations. We aim to externally validate established MPMs on a large TAVI dataset from the Netherlands Heart Registration (NHR).

Methods: We included data from NHR-patients who underwent TAVI during 2013–2017. We calculated the predicted mortalities per MPM. We assessed the predictive performance by discrimination (Area Under Receiver Operating-characteristic Curve, AU-ROC); the Area Under Precision-Recall Curve, AU-PRC; calibration (using calibration-intercept and calibration-slope); Brier Score and Brier Skill Score. We also assessed the predictive performance among subgroups: tertiles of mortality-risk for non-survivors, gender, and access-route.

Results: We included 6177 TAVI-patients with an observed early-mortality rate of 4.5% ($n = 280$). We applied seven MPMs (STS, EuroSCORE-I, EuroSCORE-II, ACC-TAVI, FRANCE-2, OBSERVANT, and German-AV) on our cohort. The highest AU-ROCs were 0.64 (95%CI 0.61–0.67) for ACC-TAVI and 0.63 (95%CI 0.60–0.67) for FRANCE-2. All MPMs had a very low AU-PRC of ≤0.09. ACC-TAVI had the best calibration-intercept and calibration-

Abbreviations: NHR, Netherlands Heart Registration ("Nederlandse Hart Registratie in Dutch"); Amsterdam UMC, Amsterdam University Medical Center - location AMC (Academic Medical Center); MPM, Mortality Prediction Models; TAVI (TAVR), Transcatheter Aortic Valve Implantation (Replacement); SAVR, Surgical Aortic Valve Replacement; EuroSCORE, European System for Cardiac Operative Risk Evaluation; STS-PRoM (STS), Society of Thoracic Surgeons Predicted Risk of Mortality; OBSERVANT, Observational Study Of Appropriateness, Efficacy, And Effectiveness of AVR-TAVR Procedures For the Treatment Of Severe Symptomatic Aortic Stenosis [14]; FRANCE-2, French Aortic National CoreValve and Edwards [15]; ACC-TAVI (ACC TVT), American College of Cardiology Transcatheter Valve Therapy; German-AV, German Aortic Valve Score; AU-ROC, Area Under the Receiver Operating-Characteristic Curve; AU-PRC, Area Under the Precision-Recall Curve; BSS, Brier Skill Score; LVEF, Left Ventricular Ejection Fraction; NYHA, New York Heart Association.

* Corresponding author at: Heart center and the department of medical informatics, Amsterdam UMC - Location AMC, University of Amsterdam, P.O. Box 22660, 1105 AZ Amsterdam, the Netherlands.
E-mail address: h.alfarra@amsterdamumc.nl (H. Al-Farra).

[1] "All authors take responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation".

https://doi.org/10.1016/j.ijcard.2020.05.039

slope. Brier Score values ranged between 0.043 and 0.063. Brier Skill Score ranged between −0.47 and 0.004. ACC-TAVI and FRANCE-2 predicted high mortality-risk better than other MPMs. ACC-TAVI outperformed other MPMs in different subgroups.

Conclusion: The ACC-TAVI model has relatively the best predictive performance. However, all models have poor predictive performance. Because of the poor discrimination, miscalibration and limited accuracy of the models there is a need to update the existing models or develop new TAVI-specific models for local populations.

1. Introduction

For a long time, surgical aortic valve replacement (SAVR) was the standard treatment for severe aortic valve stenosis. Patients with high mortality-risk were considered ineligible for SAVR and were treated medically. Transcatheter aortic valve implantation (TAVI) has emerged as an alternative procedure for medical treatment for those groups of patients [1]. Patients' selection for SAVR or TAVI depends on proper identification of the post-procedural mortality-risk. In practice, heart teams use mortality prediction models (MPM) to support their decisions on patients' selection. Many MPMs have been developed for cardiac procedures for patient selection, risk stratification and benchmarking. The European System for Cardiac Operative Risk Evaluation (EuroSCORE-I, 2003; and the newer version EuroSCORE-II, 2012) [2–4], and the Society of Thoracic Surgeons Predicted Risk of Mortality (STS-PRoM, 2009) model [5], have been widely used as MPMs for early (30-day) mortality after cardiac surgery. These cardiac surgery MPMs have been also used for TAVI-patients. The guidelines for the management of valvular heart disease (version 2012) has suggested that high mortality-risk estimates of EuroSCORE-*I* ≥20% and/or STS-PRoM >10% may serve as an appropriate indication for TAVI instead of SAVR [6]. However, EuroSCORE-I, -II and STS-PRoM (STS) were developed and internally validated for predicting early-mortality on standard cardiac-surgery patients, and not specifically for TAVI-patients. These MPMs are, therefore, missing essential TAVI-specific pre-procedural variables like access-route, balloon aortic valvuloplasty prior to TAVI and valve-type. Some studies reported that EuroSCORE overestimated the early-mortality probability after TAVI by 8% [7–10]. Both EuroSCORE-II and STS have been reported to have poor discrimination for predicting early-mortality after TAVI (with AU-ROC of 0.66 and 0.58, respectively) [11]. Also, other external-validation studies have reported their suboptimal predictive-performance (with poor AU-ROC and miscalibration) for TAVI early-mortality [12,13].

Over time, TAVI-specific early-mortality MPMs have been developed, such as OBSERVANT [14], FRANCE-2 [15], and American College of Cardiology TAVI (ACC-TAVI) [16]. The predictive performance of the TAVI-specific and the cardiac-surgery MPMs were externally validated in some studies [17–20]. In the IRRMA study [19], the TAVI-specific MPMs did not perform better than the cardiac surgery MPMs (had poor AU-ROCs and were miscalibrated). Contrariwise, TAVI-specific MPMs (ACC-TAVI and FRANCE-2) outperformed the other MPMs including the cardiac-surgery MPMs in the UK-study, although ACC-TAVI and FRANCE-2 had suboptimal predictive performance (both miscalibrated and had poor AU-ROC of 0.64 and 0.62, respectively) [20]. The most commonly used predictive performance measures in these external validation studies were discrimination (AU-ROC) and calibration [21]. Besides these predictive performance measures, there are other measures like Area under the precision-recall curve (AU-PRC) and the Brier Skill Score [22–30] that provide additional aspects on the predictive performance, to better understand the MPMs' predictive behavior.

The evidence about the external validity of the MPMs is limited and has not been investigated for TAVI-patients in the Netherlands. Therefore, we aimed in this study to externally validate and compare the existing MPMs in predicting early-mortality (30-days) after TAVI, using a large recent local dataset of TAVI-patients from the Netherlands Heart Registration (NHR) and deploy additional predictive performance measures.

2. Methods

2.1. Study design

This is a retrospective cohort study in which we used data from the Netherlands Heart Registration (NHR). Hence, instead of developing new models, we applied a set of currently used MPMs for external validation on our dataset. The study was approved by the institutional review board of the Catharina Hospital (Approval number: 2018-004). The used anonymized data conformed to the Declaration of Helsinki principles.

2.2. Selection of mortality prediction models (MPMs)

For this study we selected relevant MPMs by literature search on PubMed for published studies up to 2018. Using any of the following terms: TAVI, SAVR, mortality, early mortality, 30-day mortality, in-hospital mortality, clinical prediction models, mortality prediction model, risk score, risk stratification with any of the following terms: performance measures, discrimination, or calibration. We also searched using the following Mesh-terms: aortic valve stenosis, transcatheter aortic valve replacement, TAVR, and ROC Curve. An MPM was considered if it was published, internally validated, and used for early (30-day) mortality. MPMs with other end-points (long-term mortality) were not included in this study.

2.3. Definition of the primary outcome variable

The primary outcome variable of this study is the early post-procedural mortality, which we define as death within 30-days from the date of the TAVI procedure.

2.4. The Netherlands Heart Registration (NHR)

In the Netherlands, 16 heart centers perform TAVI for symptomatic aortic stenosis. Multi-disciplinary teams of cardiologists, surgeons and other healthcare professionals at each center decide on patients' eligibility for operation: SAVR or TAVI procedure. Data were extracted from the value-based healthcare (VBHC) program, which is a part of the Netherlands Heart Registration (NHR). In the VBHC program, which focuses on measuring and improving outcomes that matter most to patients, 22 Dutch heart centers voluntarily submit patient demographics, clinical characteristics, intervention risk factors, procedural details, mortality-status, complications and follow-up after hospital discharge [31]. In total, 13 out of 16 Dutch heart centers participated in presenting the outcomes of TAVI. Each center obtained the mortality data from the regional municipal administration registry. For this study, all data on each TAVI-procedure from January 1, 2013, to December 31, 2017 (NHR-TAVI cohort) of these 13 centers were extracted. For each patient, to be included in this study, the outcome status (early-mortality) should be available.

To obtain reliable data, the NHR has an advanced, certified data-quality control system in place, and an audit was completed by the

NHR on TAVI patient characteristics and outcomes in 2017. During that audit, NHR has examined a sample of 50 medical files among the participating centers.

2.5. Statistical analysis

For each selected MPM, the known and corresponding variables from the NHR-TAVI cohort were selected (e-component Table 2 presents the variables used from the NHR Registration to externally validate the candidate MPMs).

In the few cases in which a variable required by a model was not registered in the NHR-TAVI registration, the condition represented by the missing variable was assumed to be absent for all patients. This could theoretically induce a bias, though the same issue of non-registered variables had been described in previous external validation studies with a reported negligible bias [19,20].

For missing values of variables registered in the NHR-TAVI cohort, we assumed there were missing at random. Therefore, multiple imputations with ten imputed datasets were applied for the missing values using Multiple Imputation by Chained Equations (MICE). For each patient, the early-mortality probabilities obtained from the 10 imputed datasets were averaged. For each selected MPM, we used its logistic regression equation to predict early-mortality probability. In the equations, we used the regression coefficients as published in the original studies about the MPMs.

2.6. Predictive performance estimation

We used the following predictive performance aspects and their respective measures: discrimination by the Area Under Receiver Operating-Characteristic Curve (AU-ROC); the balance between the positive predictive value and the sensitivity by the Area Under Precision-Recall Curve (AU-PRC); calibration by the calibration-slope and -intercept; and accuracy by Brier Score and Brier Skill Score (BSS).

Discrimination measures the ability of the MPM to distinguish between survivors and non-survivors. It is quantified by the AU-ROC and is also equal to the concordance statistic (c-statistic) [24,32]. The closer the AU-ROC is to 1, the better the MPM is.

We compared AU-ROCs of various MPMs using the non-parametric method of Delong et al. [33]. Furthermore, because some variables were imputed, the AU-ROCs of the MPMs were compared before and after imputation using the method described by Venkatraman [34].

The AU-PRC summarizes the trade-off between the precision and the recall for each MPM using different probability thresholds [35]. The terms “recall” and “precision”, originating from the discipline of Information Retrieval, correspond respectively to the sensitivity and the positive predictive value. AU-PRC evaluates the fraction of true positives among the positive predictions. In a dataset where the prevalence of the event is low (imbalanced dataset), the AU-ROC does not provide insight into the balance between the recall and the precision [30,36,37]. Therefore, besides the AU-ROC, we also obtain AU-PRC. The closer the AU-PRC is to 1, the better the MPM is.

Calibration is the agreement between predicted and observed mortality rates across the full probability range. To assess calibration, we used the calibration approach formulated by Cox [38]. In this approach, an existing MPM is first used to obtain the predicted log-odds of early mortality on our external cohort. Then, using a separate logistic regression model, these log-odds themselves are used as the sole predictor of (again) early-mortality. If the original probabilities based on the existing MPM were perfect, and hence the log-odds, then the coefficients of in the linear predictor of this logistic regression model would be 0 for the intercept and 1 for the slope. Specifically, the two coefficients correspond to 1) the calibration-intercept (Calibration-in-the-large), which indicates the extent that predictions are systematically too low or too high, and 2) the calibration-slope (regression slope of the linear predictor). Good calibration is observed if the 95% confidence interval (CI) for the calibration-intercept includes 0, and the 95% CI of the calibration-slope includes 1.

For measuring the accuracy we use the Brier Score and Brier Skill Score (BSS), which summarize the deviations between the outcome and predicted probabilities at the patient level. Lower Brier Score and higher BSS indicate better accuracy. The Brier Score is the mean of the squared error and ranges between zero (perfect prediction) and one (the worst prediction) [25]. For better interpretation, the Brier Score is transformed into the BSS. The BSS measures the proportional improvement of each model's predictions over a non-informative reference MPM that simply predicts the prior probability of the event for all patients. The maximum value for BSS is 1, which indicates a perfect deterministic prediction i.e. the model could exactly predict the observed outcomes [39]. A BSS of zero means that there is no improvement compared to the predictions of the reference model. A negative BSS indicates poorer performance than the reference non-informative MPM.

For subgroup analysis in each MPM, we defined high, moderate and low mortality-risk subgroups based on the 33% and 66% probability tertiles for the non-survivors patients. The high, moderate, low subgroups of each MPM were plotted in a 100% stacked-column bar chart and compared. A good MPM would predict and allocate more cases from the non-survivors as high mortality-risk cases.

Another subgroup analysis was conducted on different subgroups defined by: age (≤75 and >75), gender (female), diabetes (yes and no), access-route (transfemoral and non-transfemoral), left-ventricular-ejection-fraction (LVEF) (<50% and ≥50%), NYHA (class-III and class-IV), and procedure-urgency (urgent, emergency and salvage).

In addition, for each MPM, we provide the density plots of the mortality probabilities for survivors and non-survivors. This chart is a variation on the histogram in which kernel smoothing is used for the plotting. A perfectly discriminating MPM will have non-overlapping density curves for survivors and non-survivors.

We use the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement for reporting [40] (E-component material; TRIPOD Checklist).

All statistical analyses were done in R software (version 3.5.1). Multiple imputations of the dataset were completed using the MICE package in R. The graphical plots were made using the ggplot2 package. The package pROC was used for constructing and testing the AU-ROCs, and the package PRROC to construct AU-PRC. A 2-tailed p-value <.05 was considered significant for all analyses.

3. Results

We found seven relevant MPMs, which is currently used for post-procedural in-hospital and early (30-day) mortality after TAVI (E-component Table 1). These MPMs were both internally validated in [3–5,14–16,41] and externally validated in [17–20] (E-component Table 4). Generally, there were three types of MPMs used for predicting early-mortality after TAVI. The first type includes cardiac surgery MPMs that were developed on standard cardiac-surgery patients. These are EuroSCORE-I, EuroSCORE-II, and STS-PRoM, with 17, 19, and 41 variables in each model, respectively. The second type includes the TAVI-specific MPMs consisting of ACC-TAVI, FRANCE-2, and OBSERVANT, with 9, 10 and 7 variables in each model, respectively. The third type includes the MPMs developed on TAVI and SAVR patients. In this category, there was one MPM, the German-AV, with 16 variables. The originally reported internally validated AU-ROCs for these MPMs ranged between 0.59 (FRANCE-2) and 0.81 (EuroSCORE-II) (E-component Table 1).

Data about 7319 patients from the NHR-TAVI registration, were obtained for this study. We exclude 1142 patients due to missing outcome mortality-status.

For this study, we obtained data of 7319 patients from the NHR-TAVI registration, to be used as an external validation dataset. We excluded 1142 patients due to missing outcome mortality-status. We included data of 6177 patients, with a 4.5% ($n = 280$) early-mortality rate.

Table 1 presents summary statistics for the baseline and the procedural characteristics of the external-validation dataset of this study (our TAVI-NHR-patients). The mean age was 80.0 years (S.D. 7.0), 51.0% of the patients were female and 56.0% had NYHA class-III and 8% NYHA class-IV. About 37.0% of the patients had an LVEF <50.0%, and 9.0% from the TAVI-procedures were Urgent. Patients with critical-preoperative-state had the highest early-mortality risk of 21.1%, Dialysis with 9.0% early-mortality risk, NYHA class-IV with 9.4%, access-route (non-transfemoral) with 8.3%, and procedure-urgency (urgent) with 6.7%.

In the NHR-TAVI cohort, the variables hypertension and atrial-fibrillation (used in STS) were not registered in the NHR-TAVI registration. Also, the variable acute-pulmonary-edema (used in FRANCE-2) was not registered (E-component Tables 2 and 3). Therefore, these variables were assumed to be absent for all patients.

Some of the MPMs' variables had missing values. Most variables (total 16 variables) had <2.0% missing values and 6 variables had >5.0% missing values (NYHA class-III, class-IV, poor-mobility, and diabetes). Details about the percentage of missing values are presented in E-component Table 3. These missing values were completed with multiple imputations. The AU-ROCs of all MPMs remained similar before and after imputations (E-component Table 5).

Table 1
Baseline characteristics of the total population and of early mortality (30-day) in the TAVI-NHR cohort before implementing multiple imputation.

Variables	Total population (Total = 6177)		Patients with early (30-day) mortality (Total = 280, early-mortality rate 4.5%)	
	Mean / Number (n)	S.D. / %	Mean / Number (n)	S.D. / %
Age (years)	80.0	±6.9	80.9	±6.6
Height (cm)	168.1	±9.4	166.6	±10.4
Weight (kg)	77.0	±15.3	73.2	±14.4
Body mass index (kg/m^2)	27.2	±4.9	26.4	±5.3
EuroSCORE I	16.3	±10.5	19.6	±13
EuroSCORE-II	6.1	±5.7	7.8	±6.8
Creatinine, μmol/L	108.2	±69.2	115.0	±67.9
eGFR mL/min/1.73 m^2	59.1	±21.3	56.0	±22.1
LVEF	50.2	±11.2	48.2	±11.8
sPAP mm Hg	31.1	±10.9	33.7	±13.1
sPAP > 60 mmHg	86	1.4	6	2.1
LVEF < 50%	2273	36.8	126	45.0
Female gender	3170	51.3	147	52.5
Chronic kidney disease	2764	44.7	112	40.0
Dialysis	87	1.5	8	2.9
Diabetes				
Diabetes, oral medication	789	13.2	29	10.4
Diabetes, insulin	420	7.0	17	6.1
Poor mobility	333	9.2	11	3.9
Chronic lung disease	1377	22.4	76	27.1
Extra-cardiac arteriopathy	1414	23.1	80	28.6
Previous cardiac surgery	1323	22.3	54	19.3
Recent myocardial infarction	119	2.0	9	3.2
Functional NYHA class				
Functional NYHA Class III	2991	56.1	140	50.0
Functional NYHA Class IV	405	7.6	38	13.6
Critical preoperative state	38	0.6	8	2.9
Procedure urgency				
Procedure urgency Elective	5415	90.8	215	76.8
Procedure urgency Urgent	536	9.0	41	14.6
Procedure urgency Emergency	15	0.3	1	0.4
Procedure weight (2 operations)	57	1.0	3	1.1
Anesthesia	3671	62.9	202	72.1
Access route				
Access route Transfemoral	4926	80.7	182	65.0
Access route Non-transfemoral	1163	19.1	96	34.3
Balloon pre-TAVI	2738	51.8	118	42.1

Values are mean ± standard deviation (S.D.) or number (n) and percentage (%).
Abbreviations: eGFR = estimated glomerular filtration rate; sPAP = systolic Pulmonary Arterial Pressure; LVEF = Left Ventricular Ejection Fraction; NYHA = New York Heart Association functional Classification; Balloon pre-TAVI = Balloon aortic valvuloplasty prior to date of TAVI.

ACC-TAVI with a predicted early-mortality of 4.4% came closest to the observed mortality (4.5%) in the NHR-TAVI cohort (Table 2). The predicted early-mortalities of the MPMs ranged from 3.4% (underestimation) for STS to 16.2% for EuroSCORE-I, which indicates an overestimation of the early-mortality risk.

The AU-ROCs ranged between 0.64 (95%CI 0.61–0.67) for ACC-TAVI to 0.58 (95%CI 0.55–0.62) for OBSERVANT (Table 2), and the highest and lowest AU-ROCs differed significantly (p-value = .007). FRANCE-2 had the second-highest discriminative ability with AU-ROC of 0.63 (95%CI 0.60–0.67) (Table 2 and Fig. 1). There was no difference between ACC-TAVI and FRANCE-2 (p-value = .54). There was no significant statistical difference between AU-ROCs of each MPM in the entire cohort before and after imputation (E-component Table 5).

For AU-PRC (trade-off between positive predictive value and sensitivity), both ACC-TAVI and FRANCE-2 had the highest AU-PRC values of 0.09.

Only for the model ACC-TAVI, the calibration-intercept 0.04 (95%CI -0.08 - 0.16) and calibration-slope 0.98 (95%CI 0.94–1.01) did not significantly deviated from their ideal values (Table 2 and Fig. 2).

In terms of accuracy, the Brier Score values were very low (<0.05) and similar for most of the MPMs of 0.04 except for EuroSCORE-I that had the worst Brier Score of 0.06. The BSS of ACC-TAVI is 0.002 and STS is 0.004 (Table 2). ACC-TAVI and FRANCE-2 predicted the high mortality-risk subgroup (among non-survivors) better than other MPMs. However, FRANCE-2 poorly classified moderate/low-risk subgroups (E-component Fig. 1).

The ACC-TAVI had the best performance (in terms of AU-ROC, AU-PRC, calibration, and accuracy) among the subgroups (Age, Gender, Diabetes, Access-Route, LVEF, NYHA-classes, and Procedure-Urgency) (E-component Table 6).

To better explain the distribution of mortality probabilities of each MPM, we graphically constructed density plots. As shown in (E-component Fig. 2), the curves for survivors and non-survivors overlapped on virtually all the probability range.

4. Discussion

This study showed that ACC-TAVI has relatively the best performance for predicting early-mortality in our TAVI-patients. However, the predictive performance of all validated MPMs in this study appears to be suboptimal. Hence these MPMs are unlikely to be useful for individual and personalized TAVI-mortality risk prediction outside their original populations. Therefore their applicability in the clinical practice (for patient selection, shared decision-making or benchmarking) is questionable in the Netherlands, and possibly in other external populations.

This study showed that the ACC-TAVI and FRANCE-2 models have the highest AU-ROC of 0.64 and 0.63, respectively, which is comparable to the AU-ROC findings of a previous external validation (UK-study) [20] (E-component Table 4). However, an AU-ROC between 0.6 and 0.7 is often regarded as poor. The originally reported AU-ROC was 0.66 for ACC-TAVI and 0.59 for FRANCE-2 [15,16].

ACC-TAVI is the only model in our study that had good calibration. This finding supports previous findings [20,42,43]. This balanced performance might be due to the similarities between the populations in these external validation studies and the development population.

ACC-TAVI and FRANCE-2 have the highest Area Under Precision-Recall Curve AU-PRC of about 0.1, but models with such low AU-PRC value are considered inadequate and have poor performance [30,37]. We could not find previous publications reporting on the AU-PRC values for ACC-TAVI and FRANCE-2. However, we believe the low AU-PRC obtained in the external validation is related to the generally low

Table 2
Predicted early-mortality, discrimination (AU-ROC, (SD) 95% CI), area under the precision-recall curve (AU-PRC), calibration-intercept (95% CI), calibration-slope (95% CI), Brier score, and Brier skill score for each MPM in the whole NHR-TAVI cohort ($N = 6177$).

Model (MPM)	Predicted early-mortality in NHR-TAVI	Discrimination AU-ROC (SD) 95% CI	AU-PRC	Calibration		Accuracy	
				Calibration-intercept (95% CI)[#]	Calibration-slope (95% CI)[#]	Brier score	Brier skill score
Surgical MPM							
STS	3.4%	0.62 (0.018) 0.58–0.65	0.08	0.31 (0.19–0.43)	0.90 (0.86–0.94)	0.043	0.004
EuroSCORE-I	16.2%	0.59 (0.018) 0.55–0.62	0.07	−1.49 (−1.61–1.37)	1.76 (1.68–1.84)	0.063	−0.47
EuroSCORE-II	5.5%	0.61 (0.017) 0.57–0.64	0.07	−0.21 (−0.34–0.09)	1.02 (0.98–1.07)	0.044	−0.03
TAVI-specific MPM							
ACC-TAVI	4.4%	0.64 (0.017) 0.61–0.67	0.09	0.04 (−0.08–0.16)	0.98 (0.94–1.01)	0.043	0.002
FRANCE-2	7.4%	0.63 (0.017) 0.60–0.67	0.09	−0.53 (−0.66–0.41)	1.21 (1.16–1.26)	0.044	−0.01
OBSERVANT	6.5%	0.58 (0.018) 0.55–0.62	0.08	−0.39 (−0.51–0.27)	1.11 (1.06–1.16)	0.044	−0.02
SAVR and TAVI MPM							
German-AV	9.0%	0.60 (0.018) 0.57–0.64	0.08	−0.76 (−0.88–0.64)	1.30 (1.25–1.36)	0.047	−0.09

Abbreviations: MPM = Mortality prediction model, SD = standard deviation, AU-ROC = area under the receiver operating characteristic curve = concordance *(c) statistic*; AU-PRC = area under precision-recall curve; SAVR = Surgical Aortic Valve Replacement.
[#] Calibration-intercepts and -slopes for each model were estimated assuming the slope(s) and intercept(s) equal to one and zero respectively. A satisfactory calibration considered if the 95%CI for the calibration-intercept and-slope included the zero and one, respectively. Bold items represent having the best predictive-performance among the other models.

prevalence of the outcome measure in the TAVI patient population and the fact that the model does have a very good discrimination ability. Outcome prevalence is hence associated with a low positive predictive value and hence a low AU-PRC.

The BSS values ranged between −0.47 and 0.004. Both ACC-TAVI and STS have BSS of just above zero, meaning there is no marked prediction improvement compared to the non-informative reference model. For the other MPMs, the BSS had negative values, indicating predictions are even poorer than the reference model (Table 2).

When analyzing the mortality-risk subgroups in early-mortality cases, the model FRANCE-2 classified 36% (102/280) of deaths as low mortality-risk patients. In contrast, ACC-TAVI classified (less and hence better) 21% (58/280) of deaths as low mortality-risk (E-component Fig. 1). This difference, which is in favor of ACC-TAVI, is due to the ability of ACC-TAVI to predict more cases of high mortality-risk from the deaths. This is likely due to the presence of three variables Acuity-Category [2–4,and] in the model ACC-TAVI, but not in FRANCE-2. These variables correspond directly or by a combination of the NHR-cohort variables Procedure-Urgency (Urgent, Emergent, and Salvage), Critical-Preoperative-Status and Recent-Myocardial-Infarction.

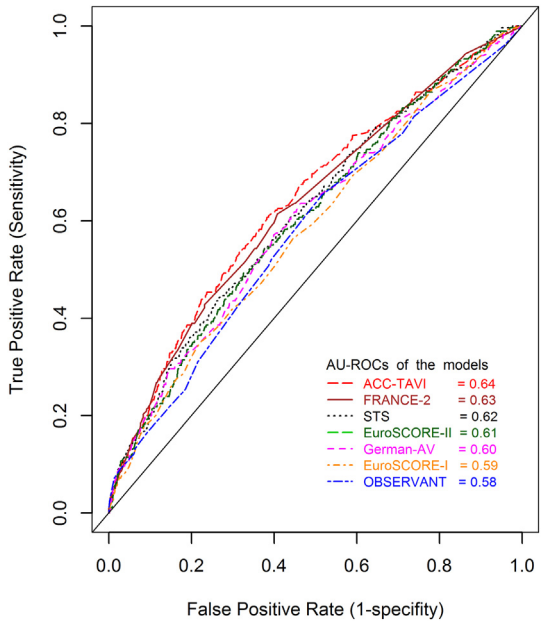

Fig. 1. The area under receiver operative curve (AU-ROC) for each of the mortality prediction models in the TAVI-NHR cohort.

In FRANCE-2, only Critical-Preoperative-Status is used for the prediction. Therefore, it seems that FRANCE-2 ignores some potential mortality variables in the NHR-TAVI-cohort.

The density plots of probabilities for survivors and non-survivors show large overlap (E-component Fig. 2), indicating the poor ability of any MPM to separate survivors from non-survivors in our population.

4.1. Strengths and limitations

The main strength of this study is the large sample size. This analysis is based on the contemporary and largest TAVI-population in the Netherlands. Nearly all heart centers in the Netherlands provided data on TAVI patients. To the best of our knowledge, this is the first study in the Netherlands that externally validated and compared the predictive-performance of seven existing MPMs on a TAVI-cohort. Besides, unlike earlier studies [17–20], we deployed additional predictive performance measures (area under precision-recall curve and Brier Skill Score).

A limitation of this study is that not all variables in the MPMs were registered in the NHR-TAVI registration for TAVI-patients. In E-component Table 2 it is visible that some variables of the STS and FRANCE-2 models are missing in the NHR registration. However, in line with other studies [19,20] we assumed that the underlying conditions (e.g. acute-pulmonary-oedema) were absent for all patients for the corresponding missing variables. In addition, we performed the analysis for FRANCE-2, one of the best performing MPMS, in which we simulated the values of the acute-pulmonary-oedema variable (the only variable missing for FRANCE-2 in our NHR dataset). In each simulation we have randomly drawn values, with a probability of 0.5 of each outcome (absent/present) and calculated the performance measures. The performance estimates and their confidence intervals were essentially the same.

Another possible limitation is the missing values of some variables (E-component Table 3). However, we implemented multiple imputations to attenuate this limitation. Missing values and multiple imputations might introduce biases. Therefore, we calculated AU-ROCs of all MPMs before and after imputations, which remained unaffected (E-component Table 5).

4.2. The implication for future work

Cardiac surgery MPMs are used routinely to justify the indication for TAVI in high mortality-risk patients. Moreover, they are used for TAVI quality control and benchmarking. Our study showed that these MPMs have poor discrimination, miscalibration and overestimated

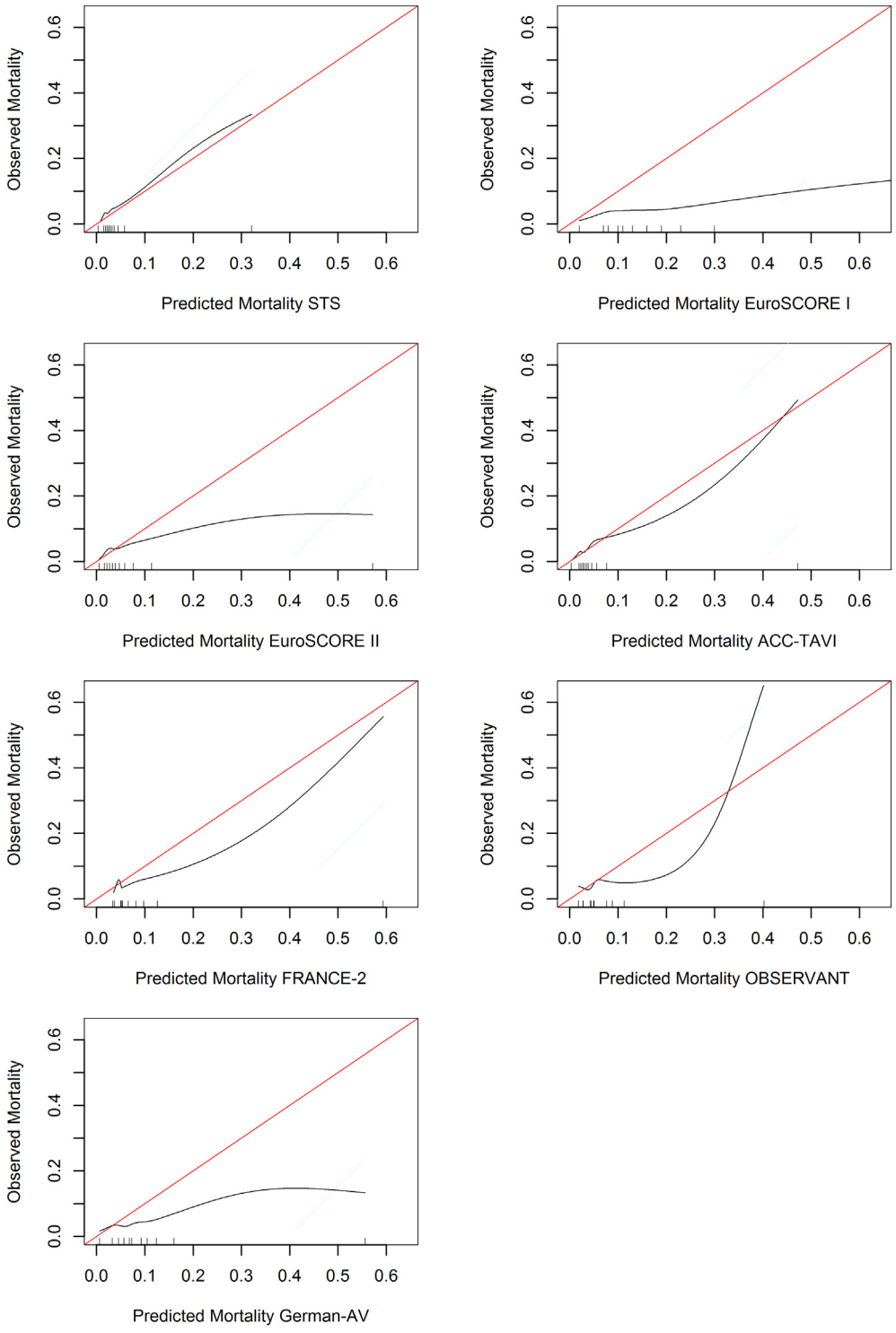

Fig. 2. Calibration plots showing the predicted vs. observed early-mortality for the mortality prediction models in the NHR-TAVI cohort. The diagonal red-line represents the perfect calibration for a perfect model (the predicted early-mortality being equal to the observed early-mortality). The black-line in each graph represents the calibration of each MPM. If the black-line is above the red-line (see STS graph), then the predicted early-mortality is lower than the observed early-mortality (i.e. underestimation). EuroSCORE I and II, FRANCE-2, German-AV overestimate the early-mortality; note the predicted early-mortality is consistently higher than the observed mortality. OBSERVANT overestimated the early-mortality in the low-risk (range x-axis from 0 to 0.33) and underestimated it for the high-risk cases (range >0.33). ACC-TAVI overestimated early mortality, but with the best calibration-on-the-large (calibration intercept) and calibration–slope (see Table 3). Despite the high density of cases in the lower range of predicted mortality, 99% of the patients have predicted values in the depicted ranges of the x-axis.

TAVI-related early-mortality, hence their use in patient selection, quality control and benchmarking is questionable.

In this study, ACC-TAVI and FRANCE-2 emerged as the best two performing MPMs in our cohort. However, they still seem relatively poor for predicting TAVI early-mortality outside their original populations.

Using a univariate analysis we found potential predictors that are not part of the set of variables in the two best performing MPMs

(ACC-TAVI and FRANCE-2). Those variables are general anesthesia (no/yes), body surface area (m2), diabetes on insulin (no/yes), LVEF (no/yes), peripheral artery disease (no/yes), age, and chronic pulmonary disease (no/yes). Including these models in a new TAVI prediction model could possibly improve the models for TAVI patients.

A new TAVI-specific MPM with better predictive performance is therefore required in order to stratify patients into high as well as moderate and low mortality risk subgroups. This is especially important as TAVI-procedures are becoming the standard therapy rather than conventional surgery.

Until a new or updated TAVI-specific MPM will be available, we encourage participating heart centers in the Netherlands to enhance the data registry.

5. Conclusion

This external validation study showed that there are large differences between the ability of the MPMs to predict early-mortality after TAVI. The ACC-TAVI model has relatively the best predictive performance. However, all studied models had poor predictive performance. Because of the poor discrimination, poor calibration and the limited accuracy of the current models, their use in clinical practice and benchmarking, at least in the Netherlands and likely in other cohorts, is questionable. This study unveiled the unmet need for developing and validation of an appropriate TAVI-specific MPM.

Funding sources

This work has no funding sources.

Declaration of Competing Interest

Hatem Al-Farra, Ameen Abu-Hanna, Bas AJM de Mol, WJ ter Burg, Saskia Houterman, José PS Henriques, and Anita CJ Ravelli declare no conflict of interest.

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

M.M. Vis, Amsterdam University Medical Centers.
J. Vos, Amphia Hospital.
J. ten Berg, St. Antonius Hospital.
W.A.L. Tonino, Catharina Hospital.
C.E. Schotborgh, HagaHospital.
V. Roolvink, Isala.
F. Porta, Leeuwarden Medical Center.
M. Stoel, Medisch Spectrum Twente.
S. Kats, Maastricht University Medical Center.
G. Amoroso, Onze Lieve Vrouwe Gasthuis.
H.W. van der Werf, University Medical Center Groningen.
P.R. Stella, University Medical Center Utrecht.
P. de Jaegere, Erasmus University Medical Center.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijcard.2020.05.039.

References

[1] A. Cribier, H. Eltchaninoff, A. Bash, N. Borenstein, C. Tron, F. Bauer, et al., Percutaneous transcatheter implantation of an aortic valve prosthesis for calcific aortic stenosis: first human case description, Circulation. 106 (24) (2002) 3006–3008.
[2] P. Michel, F. Roques, S.A. Nashef, S.P.G. Euro, Logistic or additive EuroSCORE for high-risk patients? Eur. J. Cardiothorac. Surg. 23 (5) (2003) 684–687 (discussion 7).
[3] S.A. Nashef, F. Roques, L.D. Sharples, J. Nilsson, C. Smith, A.R. Goldstone, et al., EuroSCORE II, Eur. J. Cardiothorac. Surg. 41 (4) (2012) 734–744 (discussion 44-5).
[4] F. Roques, P. Michel, A.R. Goldstone, S.A. Nashef, The logistic EuroSCORE, Eur. Heart J. 24 (9) (2003) 881–882.
[5] 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.
[6] A. Vahanian, O. Alfieri, F. Andreotti, M.J. Antunes, G. Baron-Esquivias, H. Baumgartner, et al., Guidelines on the management of valvular heart disease (version 2012): the joint task force on the management of Valvular Heart Disease of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS), Eur. J. Cardiothorac. Surg. 42 (4) (2012) S1–44.
[7] P. Genereux, S.J. Head, N.M. Van Mieghem, S. Kodali, A.J. Kirtane, K. Xu, et al., Clinical outcomes after transcatheter aortic valve replacement using valve academic research consortium definitions: a weighted meta-analysis of 3,519 patients from 16 studies, J. Am. Coll. Cardiol. 59 (25) (2012) 2317–2326.
[8] C. Tamburino, D. Capodanno, A. Ramondo, A.S. Petronio, F. Ettori, G. Santoro, et al., Incidence and predictors of early and late mortality after transcatheter aortic valve implantation in 663 patients with severe aortic stenosis, Circulation. 123 (3) (2011) 299–308.
[9] R.L. Osnabrugge, A.M. Speir, S.J. Head, C.E. Fonner, E. Fonner, A.P. Kappetein, et al., Performance of EuroSCORE II in a large US database: implications for transcatheter aortic valve implantation, Eur. J. Cardiothorac. Surg. 46 (3) (2014) 400–408 (discussion 8).
[10] J.J. Popma, G.M. Deeb, S.J. Yakubov, M. Mumtaz, H. Gada, D. O'Hair, et al., Transcatheter aortic-valve replacement with a self-expanding valve in low-risk patients, N. Engl. J. Med. 380 (18) (2019) 1706–1715.
[11] E. Durand, B. Borz, M. Godin, C. Tron, P.Y. Litzler, J.P. Bessou, et al., Performance analysis of EuroSCORE II compared to the original logistic EuroSCORE and STS scores for predicting 30-day mortality after transcatheter aortic valve replacement, Am. J. Cardiol. 111 (6) (2013) 891–897.
[12] I. Ben-Dor, M.A. Gaglia Jr., I.M. Barbash, G. Maluenda, C. Hauville, M.A. Gonzalez, et al., Comparison between Society of Thoracic Surgeons score and logistic EuroSCORE for predicting mortality in patients referred for transcatheter aortic valve implantation, Cardiovasc Revasc Med. 12 (6) (2011) 345–349.
[13] N. Piazza, P. Wenaweser, M. van Gameren, T. Pilgrim, A. Tzikas, A. Otten, et al., Relationship between the logistic EuroSCORE and the Society of Thoracic Surgeons Predicted Risk of Mortality score in patients implanted with the CoreValve ReValving system–a Bern-Rotterdam Study, Am. Heart J. 159 (2) (2010) 323–329.
[14] D. Capodanno, M. Barbanti, C. Tamburino, P. D'Errigo, M. Ranucci, G. Santoro, et al., A simple risk tool (the OBSERVANT score) for prediction of 30-day mortality after transcatheter aortic valve replacement, Am. J. Cardiol. 113 (11) (2014) 1851–1858.
[15] B. Iung, C. Laouenan, 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.
[16] 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.
[17] K. Zbronski, Z. Huczek, D. Puchta, K. Paczwa, J. Kochman, R. Wilimski, et al., Outcome prediction following transcatheter aortic valve implantation: multiple risk scores comparison, Cardiol. J. 23 (2) (2016) 169–177.
[18] V.M. Collas, C.M. Van De Heyning, B.P. Paelinck, I.E. Rodrigus, C.J. Vrints, J.M. Bosmans, Validation of transcatheter aortic valve implantation risk scores in relation to early and mid-term survival: a single-Centre study, Interact. Cardiovasc. Thorac. Surg. 22 (3) (2016) 273–279.
[19] A. Halkin, A. Steinvil, G. Witberg, A. Barsheshet, M. Barkagan, A. Assali, et al., Mortality prediction following transcatheter aortic valve replacement: a quantitative comparison of risk scores derived from populations treated with either surgical or percutaneous aortic valve replacement. The Israeli TAVR Registry Risk Model Accuracy Assessment (IRRMA) study, Int. J. Cardiol. 215 (2016) 227–231.
[20] 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.
[21] D.G. Altman, Y. Vergouwe, P. Royston, K.G. Moons, Prognosis and prognostic research: validating a prognostic model, BMJ. 338 (2009) b605.
[22] M. Tang, P. Hu, C.F. Wang, C.Q. Yu, J. Sheng, S.J. Ma, Prediction model of cardiac risk for dental extraction in elderly patients with cardiovascular diseases, Gerontology. (2019) 1–8.
[23] A. Linden, Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis, J. Eval. Clin. Pract. 12 (2) (2006) 132–139.
[24] M.S. Pepe, The Statistical Evaluation of Medical Tests for Classification and Prediction, Oxford University Press, Oxford, 2003 (07 October 2004. 320 p).
[25] G. Brier, Verification of forecasts expressed in terms of probability, Mon. Weather Rev. 78 (1950) 1–3.
[26] H. Hemingway, P. Croft, P. Perel, J.A. Hayden, K. Abrams, A. Timmis, et al., Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes, BMJ. 346 (2013), e5595. .
[27] A.D. Hingorani, D.A. Windt, R.D. Riley, K. Abrams, K.G. Moons, E.W. Steyerberg, et al., Prognosis research strategy (PROGRESS) 4: stratified medicine research, BMJ. 346 (2013), e5793. .
[28] R.D. Riley, J.A. Hayden, E.W. Steyerberg, K.G. Moons, K. Abrams, P.A. Kyzas, et al., Prognosis Research Strategy (PROGRESS) 2: prognostic factor research, PLoS Med. 10 (2) (2013), e1001380. .

[29] E.W. Steyerberg, K.G. Moons, D.A. van der Windt, J.A. Hayden, P. Perel, S. Schroter, et al., Prognosis Research Strategy (PROGRESS) 3: prognostic model research, PLoS Med. 10 (2) (2013), e1001381. .
[30] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLoS One 10 (3) (2015), e0118432. .
[31] The Netherlands Heart Registration (NHR), https://nederlandsehartregistratie.nl/ 2019 (accessed 14 December 2019).
[32] L.E. Dodd, M.S. Pepe, Partial AUC estimation and regression, Biometrics. 59 (3) (2003) 614–623.
[33] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics. 44 (3) (1988) 837–845.
[34] E.S. Venkatraman, A permutation test to compare receiver operating characteristic curves, Biometrics. 56 (4) (2000) 1134–1138.
[35] K. Boyd, K.H. Eng, C.D. Page (Eds.), Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
[36] B. Ozenne, F. Subtil, D. Maucort-Boulch, The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases, J. Clin. Epidemiol. 68 (8) (2015) 855–859.
[37] J. Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves, Proceedings of the 23rd International Conference on Machine Learning; Pittsburgh, Pennsylvania, USA, 1143874, ACM 2006, pp. 233–240.
[38] D.R. Cox, Two further applications of a model for binary regression, Oxford University Press on behalf of Biometrika Trust. 45 (1958) 562–565 (4 pages).
[39] A.H. Murphy, A new vector partition of the probability score, J Appl Meteorol, National Center for Atmospheric Research, Boulder, Colo. (1973) 595–600.
[40] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, BMJ. 350 (2015) g7594.
[41] J. Kotting, W. Schiller, A. Beckmann, E. Schafer, K. Dobler, C. Hamm, et al., German aortic valve score: a new scoring system for prediction of mortality related to aortic valve procedures in adults, Eur. J. Cardiothorac. Surg. 43 (5) (2013) 971–977.
[42] M. Arsalan, M. Weferling, F. Hecker, G. Filardo, W.K. Kim, B.D. Pollock, et al., TAVI risk scoring using established versus new scoring systems: role of the new STS/ACC model, EuroIntervention. 13 (13) (2018) 1520–1526.
[43] T. Pilgrim, A. Franzone, S. Stortecky, F. Nietlispach, A.G. Haynes, D. Tueller, et al., Predicting mortality after transcatheter aortic valve replacement: external validation of the transcatheter valve therapy registry model, Circ Cardiovasc Interv. 10 (11) (2017).