Dynamic prediction modeling of postoperative mortality among patients undergoing surgical aortic valve replacement in a statewide cohort over a 12-year period

Jackie Pollack, MSc,[a] Wei Yang, PhD,[b] Erin M. Schnellinger, PhD, MS,[c] George J. Arnaoutakis, MD,[d] Michael J. Kallan, MS,[e] and Stephen E. Kimmel, MD, MSCE[a]

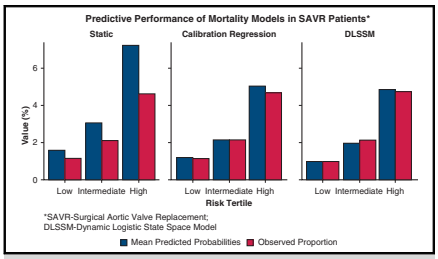
ABSTRACT

Objective: Clinical prediction models for surgical aortic valve replacement mortality, are valuable decision tools but are often limited in their ability to account for changes in medical practice, patient selection, and the risk of outcomes over time. Recent research has identified methods to update models as new data accrue, but their effect on model performance has not been rigorously tested.

Methods: The study population included 44,546 adults who underwent an isolated surgical aortic valve replacement from January 1, 1999, to December 31, 2018, statewide in Pennsylvania. After chronologically splitting the data into training and validation sets, we compared calibration, discrimination, and accuracy measures amongst a nonupdating model to 2 methods of model updating: calibration regression and the novel dynamic logistic state space model.

Results: The risk of mortality decreased significantly during the validation period ($P < .01$) and the nonupdating model demonstrated poor calibration and reduced accuracy over time. Both updating models maintained better calibration (Hosmer-Lemeshow χ^2 statistic) than the nonupdating model: nonupdating (156.5), calibration regression (4.9), and dynamic logistic state space model (8.0). Overall accuracy (Brier score) was consistently better across both updating models: dynamic logistic state space model (0.0252), calibration regression (0.0253), and nonupdating (0.0256). Discrimination improved with the dynamic logistic state space model (area under the curve, 0.696) compared with the nonupdating model (area under the curve, 0.685) and calibration regression method (area under the curve, 0.687).

Conclusions: Dynamic model updating can improve model accuracy, discrimination, and calibration. The decision as to which method to use may depend on which measure is most important in each clinical context. Because competing therapies have emerged for valve replacement models, updating may guide clinical decision making. (JTCVS Open 2023;15:94-112)



Comparison of 30-day postoperative mortality: Predicted probabilities and observed proportions by risk tertiles and updating strategies.

CENTRAL MESSAGE

Prediction models used in practice typically demonstrate poor performance over time and are infrequently updated. Dynamically updating these models over time can improve model performance.

PERSPECTIVE

Patient selection and outcomes of SAVR have continuously changed over time but existing prediction models have not. Failing to update SAVR prediction models leads to inaccurate risk assessment and risk stratification, which can lead to suboptimal treatment decisions and quality assessment. Regularly updating models can improve prediction accuracy and may lead to improved patient care.

From the [a]Department of Epidemiology, College of Public Health and Health Professions and College of Medicine, University of Florida, Gainesville, Fla; [b]Department of Biostatistics, Epidemiology, and Informatics, and [c]Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pa; [c]Research Department, United Network For Organ Sharing, Richmond, Va; and [d]Division of Cardiovascular and Thoracic Surgery, University of Texas at Austin Dell Medical School, Austin, Tex.

Supported by National Institutes of Health grant No. R01HL14129. Ms. Pollack is supported by a graduate assistantship in the Department of Epidemiology, University of Florida.

Received for publication March 13, 2023; revisions received June 7, 2023; accepted for publication June 21, 2023; available ahead of print Aug 24, 2023.

Address for reprints: Stephen E. Kimmel, MD, MSCE, Department of Epidemiology, University of Florida, 2004 Mowry Rd, PO Box 100231, Gainesville, FL 32610 (E-mail: skimmel@ufl.edu).

2666-2736

https://doi.org/10.1016/j.xjon.2023.07.011

Abbreviations and Acronyms

AS	= aortic stenosis
CPM	= clinical prediction model
BS	= Brier score
CR	= calibration regression
DPM	= dynamic prediction model
DLSSM	= dynamic-logistic state space model
EuroSCORE	= European System for Cardiac Operative Risk Evaluation
H-L	= Hosmer-Lemeshow
LASSO	= least absolute shrinkage and selection operator
MAE	= mean absolute error
PHC4	= Pennsylvania Health Care Cost Containment Council
SAVR	= surgical aortic valve replacement
STS	= Society of Thoracic Surgeons
TAVR	= transcatheter valve replacement

Aortic valvular diseases such as aortic stenosis (AS) and aortic insufficiency are leading causes of valvular morbidity and mortality in the United States and their prevalence is expected to continue rising as the population ages.[1,2] It is projected that there will be nearly 0.8 million patients with severe symptomatic AS in 2025 and 1.4 million by 2050.[3] Surgical aortic valve replacement (SAVR) is a life-saving treatment option for those with severe symptomatic AS. However, the procedure is not without substantial risk because estimates of mortality range from 1.0% to 16.4%.[4]

Well-calibrated clinical prediction models (CPMs) can serve as valuable, quick, and objective tools for risk assessment. They help determine treatment options and optimize patient care through enhanced risk communication and shared decision making.[5] CPMs typically are developed at a singular point in time in a select patient population. Although they may be validated in a separate population, they are often used for years without further updating, leading to deterioration in model performance. Even when they are updated, the process generally relies on collecting new, large samples of patients—which can take years to accrue. Moreover, with lengthy intervals between updates, models can quickly become inaccurate. This article refers to this approach as a static model approach.

The limitations of the static approach and its performance drift in the context of postoperative mortality for SAVR have been well documented in the European System for Cardiac Operative Risk Evaluation (EuroSCORE I and II) and the Society of Thoracic Surgeons (STS) models[6-13] With decreasing mortality trends for SAVR procedures, evolving care practices, and shifting patient demographics, worsening performance is a natural limitation of static models.[1,14] Further, the introduction of transcatheter valve replacement (TAVR) in 2011 as a treatment option for patients initially deemed as too high-risk for SAVR has created a dramatic shift in the patient population of SAVR procedures. Within the past few years, TAVR became the dominant treatment option for AS, even for those with intermediate mortality risk.[15-19]

Dynamic prediction models (DPMs) are a proposed solution that incorporate underlying changes over time. Although there are several time-dependent updating strategies proposed in the literature, our recent research suggests that calibration regression (CR) possesses the best set of features for dynamically updating models.[20,21] Another recently developed method, dynamic logistic state-space modeling (DLSSM), holds promise to improve on CR but has not been compared with CR in the literature. This study aimed to compare a static model's predictive ability to 2 dynamic model updating methods: CR and DLSSM in predicting 30-day postoperative mortality among patients undergoing SAVR from the state of Pennsylvania.

We hypothesize that CR and DLSSM will outperform the nonupdating approach and that DLSSM will perform the best due to its ability to examine the trend of model coefficient change over time and to potentially improve both calibration and discrimination.

MATERIALS AND METHODS

Data

The data used in this analysis are from the Pennsylvania Health Care Cost Containment Council (PHC4), which collects inpatient hospital discharge and ambulatory/outpatient procedure records from nonfederal hospitals and freestanding ambulatory surgery centers throughout Pennsylvania. Each record may document up to 18 comorbidities per patient visit. These data are collected every quarter and verified by PHC4 staff. A detailed data dictionary is available online.[22] International Classification of Disease Codes 9 and 10 were used to identify new, isolated SAVR patients along with a list of potential predictors of 30-day postoperative mortality.

Study Population

Adults aged 30 years or older who underwent an isolated SAVR from January 1, 1999, to December 31, 2018, in the state of Pennsylvania were included. This period was selected because it ensured sufficient sample size and study power, incorporated temporal changes and medical advances in the treatment of AS that likely influence one's estimated probability of survival following the surgery (eg, TAVR), and provided complete follow-up data at the time of the study. Nonresidents were excluded because out-of-state patients may not have complete follow-up information available. Patients with a history of aortic valve procedures, TAVR procedures, and concomitant cardiac procedures were excluded. We also excluded those with a primary diagnosis of shock, mechanical circulatory support, intra-aortic balloon pump, extracorporeal membrane oxygenation, cardiogenic shock, cardiac arrest, and cardiopulmonary resuscitation because they are rare events in this population and influenced the stability of the model. Lastly, we excluded those with missing mortality data (n = 29 [0.07%]) and admissions type (elective vs emergency, n = 60 [0.13%]) (See Figure 1).

The data were chronologically split into training and validation sets. The training set included 14,070 participants from 1999 to 2006. The validation set contained 30,476 participants between 2007 and 2018.

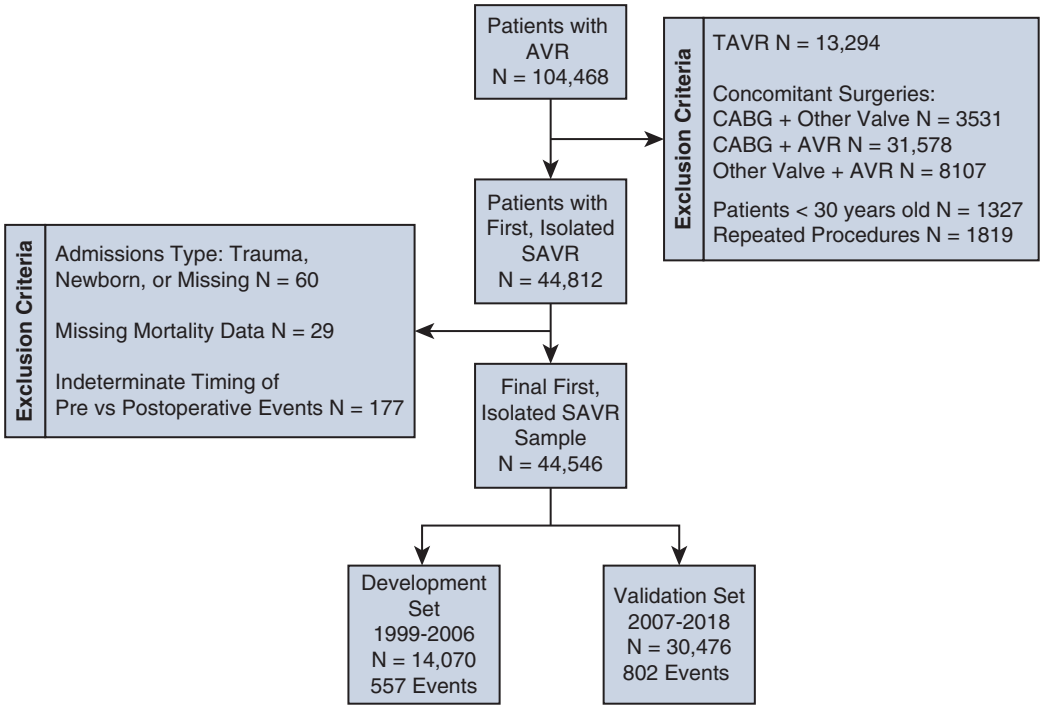


FIGURE 1. Derivation of the study sample. This figure maps the inclusion and exclusion criteria for the population used for this analysis and also shows how the data were split into the development and validation sets. *AVR*, Aortic valve replacement; *TAVR*, transcatheter valve replacement; *CABG*, coronary artery bypass graft; *SAVR*, surgical aortic valve replacement.

Outcome

The outcome was 30-day postoperative mortality, which includes in-hospital mortality and deaths within 30-days following the procedure. Mortality status was verified by PHC4 linking all patient records with the Pennsylvania Department of Health Mortality Data files.

Predictors

Initially, 40 candidate predictors were identified through literature reviews and medical expertise. We excluded 6 variables for which we could not reliably distinguish between preoperative events and perioperative/postoperative complications (arterial embolism and thrombosis, atrial fibrillation/flutter, heart block, pulmonary embolism, stroke, and ventricular fibrillation/flutter). Univariable analyses between candidate predictors and the outcome were conducted. At this stage, we found 4 variables (diabetes, depression, hypertension, and hypercholesterolemia) to have an unexplainable, significant protective association with 30-day mortality (see Table E1 for sensitivity analysis and further discussion). Because this finding is contradictory to well-established risk factors, we deemed these variables unreliable and excluded them as potential predictors. All predictors except for age were treated as binary variables. The continuous age variable was modeled linearly after examining for nonlinear associations.

Statistical Analysis

To evaluate differences in the distribution of baseline characteristics between the training and validation cohorts, the standardized mean difference was calculated for each variable. Values $\geq$0.1 were considered meaningful differences.[23]

We compared 3 approaches: the standard (static) nonupdating approach, model updating via CR, and DLSSM. Models were developed in the training set and evaluated for performance in the validation set.

Model Development and Updating

Nonupdating method. We fit logistic regression models for predicting 30-day postoperative mortality using least absolute shrinkage and selection operator (LASSO) regression for variable selection in the development cohort. The tuning parameter was selected based on minimizing model deviance and ensuring model parsimony without substantially affecting the C-statistic (see Appendix E1). We refer to this model as the LASSO model. In this approach, the LASSO model is unchanged in the validation set.

Logistic CR. Logistic CR starts with the LASSO model (ie, initial model) in the static method and annually updates the model coefficients each year within the validation set. Beginning in 2007, a logistic regression model is fit with the predicted probability of mortality (in log odds scale) estimated from the initial model as the only covariate. The coefficients from the logistic CR are subsequently used to update the predicted probabilities estimated from the initial model (see Figure E1 and Appendix E1 for further details).

DLSSM. For the DLSSM model, we used the R DLSSM package (RStudio, PBC, 2023) to fit a model using the same covariates as the LASSO model. DLSSM can examine the trend of model coefficient change over time, which is modeled using smoothing splines. The corresponding smoothing parameter is chosen by maximum likelihood. We fit 8 DLSSM models within the training set, each of which allows the coefficient for 1 of the 8 covariates to change over time in addition to the a priori specified time-varying intercept. A variable is considered to have meaningfully changed over time when the 95% confidence bands excludes the initial point estimate. We found no evidence that the coefficient for any of the 8 covariates is time-varying (see Figure E2). Therefore, the final DLSSM model from the training set included only the time-varying intercept; the other coefficients remain time-invariant (see Figure E3).

In the validation set, DLSSM continuously updates model parameters every year. Unlike CR, which rescales the predicted probability using only the recalibrated intercept and slope, DLSSM is more flexible by

updating each model coefficient individually. For a more detailed explanation refer to Jiang and colleagues[24] and Appendix E1.

Model assessment. Both calibration and discrimination are important measures of prediction performance. Calibration refers to the differences between observed and predicted probabilities of the outcome. We assessed calibration through the Hosmer-Lemeshow (H-L) statistic, calibration plots of predicted versus observed mortality, and by the calibration intercept and slope.[25] Discrimination measures how well models can differentiate between those who did and did not develop the outcome and was measured with the C-statistic. Overall accuracy was measured by the Brier score (BS) and mean absolute error (MAE).[26]

Data analyses and graphical outputs were performed using R version 4.1.2 (R Foundation for Statistical Computing). The study was reviewed by the University of Florida's Internal Review Board and received authorization as nonhuman subject research and deemed exempt (entry ID 17591; February 2, 2023).

RESULTS

Participants

Figure 1 shows the derivation of the study cohort. A total of 44,546 SAVR procedures were included in this analysis. The LASSO and DLSSM models were developed using participants from 1999 to 2006 (N = 14,070; 557 deaths). The validation set included patients undergoing SAVR from 2007 to 2018 (N = 30,476; 802 deaths). Each year within the validation set had approximately 2000 to 3000 participants. The annual risk of mortality ranged from approximately 2.0% to 4.5% and decreased significantly ($P < .001$) over the study period (Figure 2).

The distribution and characteristics of the study population stratified by training and validation set are presented in Table 1. The average age of patients undergoing SAVR in the development set was 66.6 ± 13.3 years and 67.4 ± 12.3 years in the validation set. Within both sets, patients undergoing SAVR were predominately men and undergoing an elective (nonemergency) procedure. In general, patients in the validation set had more comorbidities than those in the training data (standardized mean difference >0.1). Only chronic kidney disease stage 5 and an emergency admission were more common in the training set.

Model Specification

The LASSO and DLSSM models specified 8 covariates with similar values between the 2 models. The model covariates and performance are shown in Table 2.

Comparison of Updating Strategies

Calibration. The static model demonstrated worse calibration, overpredicting the probability of mortality in the validation cohort with a H-L statistic = 156.490 ($P < .001$) (Figure 3 and Table 3). The H-L statistic also predominantly increased each year (Table E2).

The CR and DLSSM models demonstrated better calibration and similar performance in the validation cohort (Table 3 and Figure 3). Within each year of updating (Table E2), the updating methods demonstrated better calibration (H-L statistic) than the static method, as reflected in the calibration plots (Figure E4). Consistent with these results, the intercept (ie, calibration-in-the-large) was closer to zero (better calibration) and the slope was closer to 1 for both updating methods compared with the static method (Table 3). Although there was more variability in the year-to-year evaluation of the intercepts and slopes across the 3 models (Table E2), overall, the updating models showed better performance compared with the static model.

The static model consistently overpredicted the risk of 30-day postoperative mortality. For example, a 64-year-old patient who survived and was admitted for an emergency procedure with diagnoses of aortic aneurysm/dissection and heart failure had a predicted probability consistently around 10% under the static model. Both updating methods

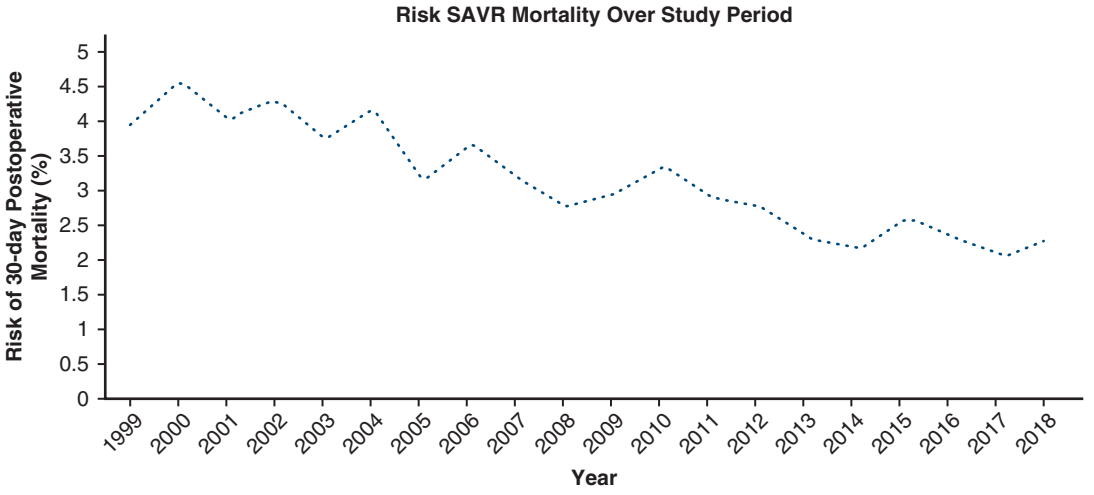


FIGURE 2. Risk of 30-day postoperative mortality after surgical aortic valve replacement (*SAVR*) in the patient population. This figure plots the risk of mortality among patients undergoing SAVR over time throughout the study period.

TABLE 1. Characteristics of patient population stratified by development and validation cohorts

Characteristic	Participants		
	Development 1999-2006 (n = 14,070)	Validation 2007-2018 (n = 30,476)	SMD*
Mean age (y)	66.6 ± 13.3	67.4 ± 12.3	0.068
Acute myocardial infarction as primary diagnosis	169 (1.2)	187 (0.6)	0.062
Admission type: Emergency	4826 (34.3)	7822 (25.7)	**0.189**
Alzheimer/dementia	24 (0.2)	172 (0.6)	0.065
Anemia	2615 (18.6)	14,791 (48.5)	**0.669**
Aortic aneurysm and/or dissection	2150 (15.3)	6351 (20.8)	**0.145**
Aortic root surgery: concomitant	1715 (12.2)	5065 (16.6)	**0.126**
Asthma	612 (4.3)	1736 (5.7)	0.062
Cardiomyopathy	864 (6.1)	3043 (10.0)	**0.142**
Cachexia	12 (0.1)	47 (0.2)	0.020
Chronic pericardial disease	165 (1.2)	379 (1.2)	0.006
Chronic kidney disease stage 1-4	254 (1.8)	3286 (10.8)	**0.376**
Chronic kidney disease stage 5+	468 (3.3)	512 (1.7)	**0.106**
Chronic liver disease	169 (1.2)	678 (2.2)	0.079
Chronic obstructive pulmonary disease	2107 (15.0)	4000 (13.1)	0.053
Other chronic lung diseases	145 (1.0)	396 (1.3)	0.025
Coronary artery disease	4017 (28.6)	10,278 (33.7)	**0.112**
Endocarditis	465 (3.3)	1075 (3.5)	0.012
Excision of other lesion/heart tissue same day	325 (2.3)	1754 (5.8)	**0.176**
Heart failure	4834 (34.4)	10,307 (33.8)	0.011
History of chronic steroid use	13 (0.1)	199 (0.7)	0.092
Lupus	65 (0.5)	141 (0.5)	<0.001
Oxygen dependence therapy	26 (0.2)	518 (1.7)	**0.157**
Obstructive sleep apnea	51 (0.4)	2935 (9.6)	**0.435**
Parkinsonism	46 (0.3)	137 (0.4)	0.020
Peripheral arterial disease	2160 (15.4)	6477 (21.3)	**0.153**
Percutaneous transluminal coronary angioplasty/stent	375 (2.7)	1379 (4.5)	**0.100**
Pulmonary hypertension	405 (2.9)	855 (2.8)	0.004
Rheumatoid arthritis	124 (0.9)	340 (1.1)	0.024
Sex: Female	6067 (43.1)	11,956 (39.2)	0.079

Values are presented as mean ± SD or n (%). *SMD*, Standardized mean difference. *Values ≥ 0.1 were considered meaningful differences between the development and validation cohorts and appear in bold font.

yielded more appropriate, lower predicted probabilities consistent with the decreasing risk of the procedure over time and the survival of the patient. From the beginning of the validation period, DLSSM generated a slightly lower predicted probability of 8.6% and by 2018, after 11 years of updating, the predicted risk was nearly half that of the static model at 5.5%. As expected, DLSSM also demonstrated a smoother change in the predicted risk over time. The CR method also demonstrated a decreasing trend over time (from 10.2% down to 4.1%), though the trend was not as smooth (see Appendix E1 and Figure E5).

Discrimination (C-statistic). CR does not change the rank order of predicted risk, so as expected, it did not change the year-to-year C-statistic in the validation data (Table E2). However, when examined across years, CR yielded marginally higher C-statistics (0.687) than nonupdating (0.685) as ranking can change when combining data across updating intervals (Table 3). DLSSM demonstrated the best discrimination (C-statistic = 0.696) (Table 3), and yearly comparisons show that the DLSSM model had better areas under the curve in most years compared with CR or nonupdating models (see Appendix E1 and Table E2).

TABLE 2. Specification of logistic regression models developed from the training data (years: 1999-2006; N = 14,070)

Variable	LASSO-derived models		DLSSM-derived models	
	β Coefficient	Odds ratio (95% CI)*	β coefficient	Odds ratio (95% CI)*
Age	0.04	1.04 (1.03-1.05)	0.04	1.04 (1.03-1.05)
Acute myocardial infarction as primary diagnosis	0.82	2.27 (1.37-3.60)	0.83	2.29 (1.82-2.77)
Admission type: Elective vs emergency	0.57	1.77 (1.47-2.12)	0.55	1.73 (1.42-2.10)
Aortic aneurysm and/or dissection	0.88	2.42 (1.93-3.01)	0.84	2.31 (1.82-2.94)
Chronic kidney disease stage 5+	0.54	1.72 (1.20-2.40)	0.63	1.88 (1.30-2.73)
Chronic liver disease/cirrhosis	1.22	3.40 (1.99-5.50)	1.24	3.46 (1.99-5.94)
Endocarditis	1.06	2.88 (1.97-4.10)	1.06	2.88 (2.00-4.19)
Heart failure	0.54	1.71 (1.43-2.05)	0.52	1.68 (1.38-2.04)
Model performance				
Hosmer-Lemeshow statistic	9.36†		11.703†	
C-Statistic	0.724		0.750	
Brier score	0.037		0.034	
Mean absolute error	0.074		0.066	

The LASSO derived model was used for the nonupdating and calibration regression models. The DLSSM model uses the same variables identified by the LASSO and allows the intercept to change over time. *LASSO*, Least absolute shrinkage and selection operator; *DLSSM*, dynamic-logistic state space model. *The intercept coefficient for the LASSO model is −6.73. The intercept for the DLSSM model varied over time. For more details refer to Figure E2. †$P > .15$.

Overall accuracy. Although differences between the BS are difficult to interpret, the BS was consistently better (lower) in the updating models, both within yearly comparisons and overall (Tables 3 and E2). DLSSM had the best BS (0.0252). The MAE demonstrated similar results, with lower MAEs in the updating versus static strategies (Table 3). Again, DLSSM was the best model with the lowest MAE (0.050) compared with CR (MAE = 0.052) and nonupdating (MAE = 0.063). The improvement in MAE progressively widened among the strategies with each successive update (see Appendix E1 and Table E2).

DISCUSSION

This study evaluated 2 methods for dynamically updating a CPM in a population of patients undergoing isolated SAVR over a period of 12 years and compared these methods with the more typical nonupdating approach and with each other. The primary implication of our analyses is that regularly updating a CPM is superior to nonupdating. The nonupdating model consistently overpredicted the risk of mortality and tended to worsen in performance over time.

Among the updating strategies, DLSSM was marginally but almost consistently better than CR. Although the differences are not large, an advantage of DLSSM is that it allows each coefficient to be updated independently, altering the rank order of the predicted probabilities, and thereby improving model discrimination. In comparison, CR does not change the rank order of predicted probabilities. Therefore, in any 1 year, updating through CR does not alter model discrimination. Discrimination is more important

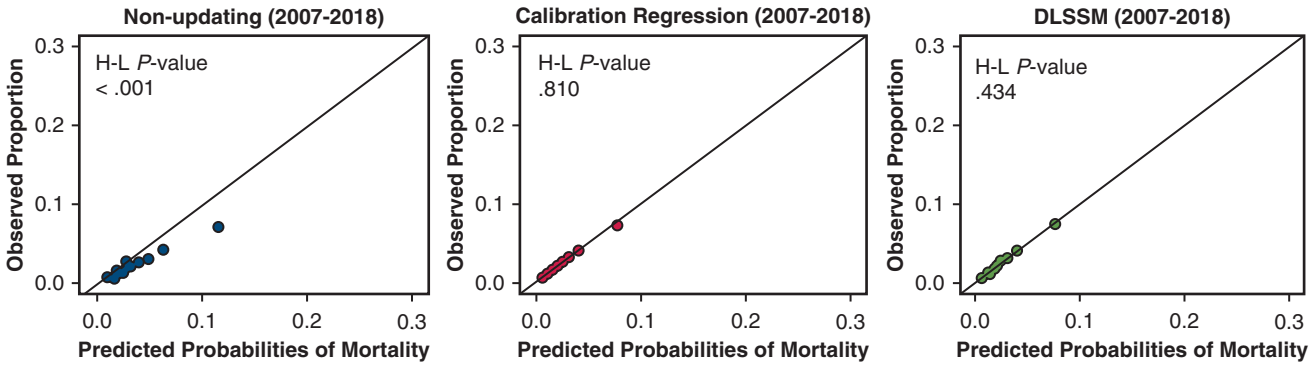


FIGURE 3. Calibration plots of updating strategies. This figure shows the predicted (*x*-axis) by observed (*y*-axis) mortality by risk deciles in the validation sample. *DLSSM*, Dynamic logistic state space model; *H-L*, Hosmer-Lemeshow χ^2 statistic with the corresponding *P*-value.

TABLE 3. Comparison of model performance in the validation set

2007-2018	Model		
Performance metric	Nonupdating	Calibration regression	DLSSM
Calibration			
Hosmer-Lemeshow	156.490	4.491	7.999
Intercept*	−0.737	−0.273	−0.202
Slope†	0.897	0.933	0.935
Discrimination			
C-Statistic	0.685	0.687	0.696
Overall accuracy			
Brier score	0.0256	0.0253	0.0252
Mean absolute error	0.063	0.052	0.050

DLSSM, Dynamic-logistic state space model. *The intercept measures calibration-in-the-large and refers to the difference between mean expected and mean observed mortality. Values closer to 0 are better with 0 indicating a perfectly calibrated model. †A slope of 1 is a perfectly calibrated model.

when assessing individual risk while improving calibration may have greater influence on risk-adjustment and institutional comparisons. How DPM might alter measures of center or operator performance is an important question. However, the methods to make these comparisons in a dynamically updating approach are still being developed and is an important topic for future research. Because the purpose of our analysis was to evaluate the summative performance of DPM, measures of calibration, discrimination, and accuracy were considered; thus, overall, DLSSM was the best method in our setting.

The findings of our analysis are similar to and expand upon our previous work of predicting 1-year survival after lung transplant, in which DPMs outperformed nonupdating.[20] Other studies have also documented dynamically updating models, but they have primarily focused on the estimation of model coefficients rather than prediction accuracy. McCormick and colleagues[27] applied a modeling strategy similar to DLSSM among children receiving either laparoscopic or open appendectomies. Their approach did not model the smoothing trend of model coefficient change as DLSSM and the work focused primarily on the relationship between covariates and procedure type, rather than prediction.[27] Hickey and colleagues[28] compared periodic refitting at varying 1- and 2-year intervals by updating strategy for in-hospital mortality following cardiac surgery, but they focused on the changing model coefficients over time and did not differentiate between types of cardiac surgeries.

There are several prediction models for SAVR mortality, but we are unaware of any models that systematically update on a regular basis. In North America, the STS Adult Cardiac Surgery Database is used to develop prediction models for major cardiac procedures, including isolated SAVR, to estimate a patient's probability of mortality, among other outcomes. The STS database has been the prototype for other surgical disciplines and has enabled risk stratification for individual patients, facilitating both individual patient counseling as well as clinical research trial design. Furthermore, the influence on improved quality that has emerged from STS database efforts must be emphasized. However, these models are based on data that are several years old, before development and the process of training, validating, and deploying the model can be substantial.[6-8] The STS regularly updates the database and applies a year and procedure type-specific correction factor to its institutional reports annually,[8] but the model itself is not updated annually. As a result, the online risk calculator, a decision support tool used by providers throughout the United States, is used until a new model is developed.[8] Further, the annual correction done in STS only recalibrates the models so that the observed-to-expected mortality ratio is equal to the overall event rate for that calendar year.[8] This method only updates the intercept as opposed to CR, which updates the intercept and slope.

In 2015, Vassileva and colleagues[29] compared the 2008 STS predicted risk of operative mortality online calculator for aortic valve replacement patients following a previous coronary artery bypass grafting to a cohort-specific recalibrated risk model. The online risk calculator overestimated the risk of operative mortality, demonstrating a need to move away from static approaches and towards more frequent updating, especially given some centers' reliance on online risk calculation for individual treatment decisions.[29] The latest 2018 STS predicted risk of operative mortality (online calculator version 4.2) showed good calibration and moderate discriminatory ability at the time, but runs the risk of becoming outdated because the model was developed from data 9 to 12 years ago (between 2011 and 2014).[30]

Another prominent example in cardiac surgery is the EuroSCORE model, which was published in 1999 and tended to overestimate mortality for low-risk patients undergoing SAVR and underestimate mortality for high-risk patients in other cardiac surgeries over time.[9-11] The model was subsequently updated in 2012 (EuroSCORE II) but has not been updated since.[12] Emerging evidence

suggests that EuroSCORE II performance may be deteriorating for some patients undergoing SAVR, particularly those aged 75 years or older.[13]

Our study is not without limitations. Our data did not demonstrate any meaningful time-varying covariates in the training set that could be used to inform predictions in the validation set. More studies are needed to examine how DLSSM would perform in other dynamic prediction settings with time-varying coefficients that might inform future updates.

Several other time-dependent updating strategies proposed in the literature such as calibration-in-the-large, the closed testing procedure, and model revision were not evaluated here.[31-33] In our previous work with the Lung Allocation Score, we found that CR required minimal data, led to more consistent improvements, and exhibited less variability over time, making it more suitable for adapting to variations in the prevalence of a binary outcome compared with other updating strategies.[20,21] We therefore chose to validate this method here. We were not able to incorporate some common measures used in prominent CPMs of mortality for patients undergoing SAVR in our initial models, such as hypertension, hypercholesterolemia, diabetes, or laboratory measurements. This may limit the generalizability of our findings but because our goal was to assess the effectiveness of dynamic updating, not to develop a new model for clinical practice, our inferences are still valid. The generalizability of our finding in other datasets with more granular data, such as STS, will need to be addressed in future studies. Still, the variables we were able to incorporate have documented prognostic importance, and our models had moderate discriminatory ability.

Strengths

This study has several strengths. First, this is, to our knowledge, the first study to empirically evaluate the performance of DLSSM compared with CR (ie, a more conventional updating strategy) in a large, statewide sample of patients undergoing SAVR. Second, the PHC4 is a statewide agency that includes data from all nonfederal hospitals in the state, ensuring a mix of complicated and less complicated procedures across a broad range of practices. This case mix helps mitigate selection bias in terms of participants. Third, our sample is sufficiently large to satisfy a minimum of 10 events per predictor in our training set, allowing for a more accurate estimation of the regression coefficients in our models. Our large sample also permitted yearly updating with approximately 50 events per year, the same number that we used to demonstrate benefit in the Lung Allocation Score.[20] Lastly, our outcome measure of 30-day postoperative mortality is robust as we restricted our sample to Pennsylvania residents and linked patient records with death certificates in Pennsylvania.

CONCLUSIONS

Our study adds insight into the reliability of dynamic updating. Prior studies have not examined the performance of repeatedly updating models over time. Our findings suggest that DPMs are superior to static models and that updates can be done with standard computing resources. In our study, DLSSM was the optimal updating strategy because it has the advantage of being able to improve both discrimination and calibration, whereas CR can only improve calibration. The decision as to which updating strategy to use may be dependent on the clinical context and logistical considerations such as the availability of data (both in terms of size and frequency of collection), computational resources, and which performance metrics one wants to optimize. In the current era of rapidly evolving transcatheter strategies for valvular interventions, dynamically updating CPMs can guide clinicians toward the best valve replacement option.

Conflict of Interest Statement

The authors reported no conflicts of interest.

The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

References

1. Bevan GH, Zidar DA, Josephson RA, Al-Kindi SG. Mortality due to aortic stenosis in the United States, 2008-2017. *JAMA*. 2019;321:2236-8.
2. Coffey S, d'Arcy JL, Loudon MA, Mant D, Farmer AJ, Prendergast BD, OxVALVE-PCS group. The OxVALVE population cohort study (OxVALVE-PCS)—population screening for undiagnosed valvular heart disease in the elderly: study design and objectives. *Open Heart*. 2014;1:e000043.
3. Osnabrugge RLJ, Mylotte D, Head SJ, Van Mieghem NM, Nkomo VT, LeReun CM, et al. Aortic stenosis in the elderly: disease prevalence and number of candidates for transcatheter aortic valve replacement: a meta-analysis and modeling study. *J Am Coll Cardiol*. 2013;62:1002-12.
4. Thiagarajan K, Jeevanantham V, Van Ham R, Gleason TG, Badhwar V, Chang Y, et al. Perioperative stroke and mortality after surgical aortic valve replacement: a meta-analysis. *Neurology*. 2017;22:227-33.
5. Wessler BS, Lai Yh L, Kramer W, Cangelosi M, Raman G, Lutz JS, et al. Clinical prediction models for cardiovascular disease: Tufts predictive analytics and comparative effectiveness clinical prediction model database. *Circ Cardiovasc Qual Outcome*. 2015;8:368-75.
6. O'Brien SM, Shahian DM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. *Ann Thorac Surg*. 2009;88:S23-42.
7. Shahian DM, Jacobs JP, Badhwar V, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2018 adult cardiac surgery risk models: Part 1—background, design considerations, and model development. *Ann Thorac Surg*. 2018;105:1411-8.
8. Jin R, Wang M, Grunkemeier GL, Furnary AP. Calibration factors for STS risk model predictions: why, how and when they are used. *Ann Thorac Surg*. 2022;113:386-91.
9. Nashef SA, Roques F, Michel P, Gaudeucheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardio Thorac Surg*. 1999;16:9-13.
10. Wendt D, Osswald BR, Kayser K, Thielmann M, Tossios P, Massoudy P, et al. Society of Thoracic Surgeons score is superior to the EuroSCORE determining mortality in high risk patients undergoing isolated aortic valve replacement. *Ann Thorac Surg*. 2009;88:468-74.

11. Sabrina S, Groenwold RHH, de Heer F, Bots ML, van der Graaf Y, van Herwerden LA. Performance of the original EuroSCORE. *Eur J Cardio Thorac Surg*. 2012;41:746-54.
12. Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. EuroSCORE II. *Eur J Cardio Thorac Surg*. 2012;41:734-45.
13. Taleb Bendiab T, Brusset A, Estagnasié P, Squara P, Nguyen LS. Performance of EuroSCORE II and Society of Thoracic Surgeons risk scores in elderly patients undergoing aortic valve replacement surgery. *Arch Cardiovasc Dis*. 2021;114:474-81.
14. Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K, et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur J Cardio Thorac Surg*. 2013;43:1146-52.
15. Glenn K. Report finds TAVR is dominant form of aortic valve replacement, outcomes steadily improving in the United States. Accessed August 2, 2023. https://www.acc.org/about-acc/press-releases/2020/11/16/18/53/report-finds-tavr-is-dominant-form-of-aortic-valve-replacement-outcomes-steadily-improving-in-the-us#:~:text=Researchers%20also%20found%20a%20steady,to%20the%20current%20TAVR%20period
16. Kim KM, Shannon F, Paone G, Lall S, Batra S, Boeve T, et al. Evolving trends in aortic valve replacement: a statewide experience. *J Card Surg*. 2018;33:424-30.
17. Bowdish ME, D'Agostino RS, Thourani VH, Schwann TA, Krohn C, Desai N, et al. STS adult cardiac surgery database: 2021 update on outcomes, quality, and research. *Ann Thorac Surg*. 2021;111:1770-80.
18. Carroll JD, Mack MJ, Vemulapalli S, Herrmann HC, Gleason TG, Hanzel G, et al. STS-ACC TVT Registry of transcatheter aortic valve replacement. *Ann Thorac Surg*. 2021;111:701-22.
19. Young MN, Kearing S, Malenka D, Goodney PP, Skinner J, Iribarne A. Geographic and demographic variability in transcatheter aortic valve replacement dispersion in the United States. *J Am Heart Assoc*. 2021;10:e019588.
20. Schnellinger EM, Yang W, Kimmel SE. Comparison of dynamic updating strategies for clinical prediction models. *Diagn Progn Res*. 2021;5:20.
21. Schnellinger EM, Yang W, Harhay MO, Kimmel SE. A comparison of methods to detect changes in prediction models. *Methods Inf Med*. 2022;61:19-28.
22. PHC4.org. Inpatient discharge data file and supporting documentation file layouts (UB04), 1990-present. 1990. Accessed May 9, 2023. https://www.phc4.org/services/datarequests/docs/specialrequests1990-Current_inpatient.pdf
23. Normand S-LT, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, et al. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *J Clin Epidemiol*. 2001;54:387-98.
24. Jiang J, Yang W, Schnellinger EM, Kimmel SE, Guo W. Dynamic logistic state space prediction model for clinical decision making. *Biometrics*. 2023;79:73-85.
25. Cox DR. Two further applications of a model for binary regression. *Biometrika*. 1958;45:562-5.
26. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2010;21:128-38.
27. McCormick TH, Raftery AE, Madigan D, Burd RS. Dynamic logistic regression and dynamic model averaging for binary classification. *Biometrics*. 2012;68:23-30.
28. Hickey GL, Grant SW, Caiado C, Kendall S, Dunning J, Poullis M, et al. Dynamic prediction modeling approaches for cardiac surgery. *Circ Cardiovasc Qual Outcomes*. 2013;6:649-58.
29. Vassileva CM, Aranki S, Brennan JM, Kaneko T, He M, Gammie JS, et al. Evaluation of the Society of Thoracic Surgeons online risk calculator for assessment of risk in patients presenting for aortic valve replacement after prior coronary artery bypass graft: an analysis using the STS adult cardiac surgery database. *Ann Thorac Surg*. 2015;100:2109-16.
30. Shahian DM, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC Jr, et al. The Society of Thoracic Surgeons 2018 adult cardiac surgery risk models: part 2—statistical methods and results. *Ann Thorac Surg*. 2018;105:1419-28.
31. Vergouwe Y, Nieboer D, Oostenbrink R, Debray TPA, Murray GD, Kattan MW, et al. A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med*. 2017;36:4529-39.
32. Jenkins DA, Sperrin M, Martin GP, Peek N. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagnos Prognos Res*. 2018;2:23.
33. Su T-L, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods for clinical prediction models. *Stat Methods Med Res*. 2018;27:185-97.

Key Words: clinical prediction model, model updating, model recalibration, surgical aortic valve replacement, dynamic logistic state space model

APPENDIX E1. SUPPLEMENTAL MATERIAL

Data Dictionary

A detailed data dictionary of PHC4 data is available online[E1] with the hyperlink below.

https://www.phc4.org/services/datarequests/docs/specialrequests1990-Current_inpatient.pdf.

Sensitivity Analysis

Univariable analyses between candidate predictors and our primary outcome were conducted. At this stage, we found 4 variables (diabetes, depression, hypertension, and hypercholesterolemia) to have a significant protective association with 30-day mortality.

Sensitivity analyses were done to determine if these comorbidities may be undercoded/not included due to the limited number of fields available per patient. Each patient can have up to 18 diagnosis codes per visit, so we looked at the proportion of fields for those with and without diabetes, depression, hypertension, and hypercholesterolemia. We found no difference in the proportion of codes used between patients with and without these conditions. We also searched the records to determine if any of these 4 conditions may have been missed by being listed as a complication of an existing condition. For example, if an individual had an International Classification of Diseases 10 code of I13- "Hypertensive heart and chronic kidney disease with heart failure and stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease" but was not accounted for in the hypertensive population. After conducting the sensitivity analysis and assessing the relationships after multivariable adjustment the protective association remained.

We also considered the influence of retaining these variables and fit a model using least absolute shrinkage and selection operator (LASSO), which selected diabetes, hypertension, and hypercholesterolemia despite their protective association with mortality. However, the C-statistic for this model was 0.75 (which included the same 8 covariates from the baseline plus diabetes, hypertension, and hypercholesterolemia) compared with 0.72 from our baseline model without them. Because there was a nominal gain in discrimination with these variables, the relationship was biologically implausible, and the associations contradictory to the literature, we deemed these variables unreliable and excluded them as potential risk factors in our analysis. It is important to note that although we did not include these variables in our models, our goal was to assess the effectiveness of dynamic updating, not to develop a new model for clinical practice. The variables we were able to incorporate in our models have documented prognostic importance in other models.[E2,E3]

Finally, these 4 diagnosis codes have had limited and varied importance in other models for aortic valve postoperative mortality developed by the Pennsylvania Health Care Cost Containment Council (PHC4). To our knowledge based on the public technical reports, hypertension is only included in one PHC4 model (2005-2006) but not others. The codes for depression, hypercholesterolemia, and diabetes have not been incorporated into prior PHC4 models as of prognostic importance. Although a separate variable for "current insulin use" has been used in some models (2005-2006,2006-2007, and 2007-2008), it has not been included since (we do not have access to this code in the PHC4 data available to us).[E4-E6]

Model Development

We fit logistic regression models for predicting 30-day postoperative mortality using LASSO regression for variable selection in the development cohort. The tuning parameter was selected based on minimizing model deviance. Two lambda values corresponding to the minimum and 1 SE above the minimum were considered, where the latter provides a more parsimonious model. We compared the performance of the 2 logistic regression models. The final model from the development phase was selected based on balancing model parsimony and prediction performance measured by the area under the receiver operating-characteristic curve (ie, C-statistic). The C-statistics were similar for the two models (0.74 for the minimum value of lambda and 0.72 for 1 SE above), so the more parsimonious LASSO model was chosen to test the updating strategies in the validation cohort (Table 2).

Methodological Overview of Calibration Regression

In calibration regression, a new model is fit with the linear predictor (lp) from the original model and the intercept as the only 2 covariates. The linear predictor is the model intercept plus the coefficients from the original model multiplied by the values in the new setting. This serves as the adjustment factor that rescales the slope. The updated intercept (α) for the new model is obtained by adding the intercept of the original model to the intercept of the new model: $\mathrm{lp} = \alpha_{new+}\ \alpha_{\mathrm{original}+}\ (\beta_{\mathrm{original}} * \mathrm{x}_{\mathrm{i}\ldots\mathrm{new}})$. The updated slope ($\beta$) is the result of multiplying the slope from the new model by the slope of the original model (β_new = β_original *β_new_model). Subsequent updates follow the same procedure, using the linear predictor from the most recently updated model.

Overview of the Dynamic-logistic State Space Model

In contrast to calibration regression, the dynamic-logistic state space model (DLSSM) can provide more timely and

accurate predictions by incorporating new data and information as they become available, allowing for more flexibility in responding to advances in the field. The primary advantage of the DLSSM method over any of the calibration measures, like the approach used in the Society of Thoracic Surgeons database, is that DLSSM may improve both discrimination and calibration, whereas other methods only improve calibration. DLSSM was updated yearly in our study (based on the primary outcome event rate), but it can also be updated as frequently as statistically possible, which is another advantage over other methods. In addition, DLSSM updates the coefficients through smoothing splines, providing a more stable and detailed modeling process, and providing valuable insights into the factors that influence outcomes so that large sudden shifts in a patient's predicted probability may be avoided.

E-References

E1. PHC4.org. Inpatient discharge data file and supporting documentation file layouts (UB04), 1990-Present; 1990. Accessed May 30, 2023. https://www.phc4.org/services/datarequests/docs/specialrequests1990-Current_inpatient.pdf

E2. O'Brien SM, Shahian DM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. *Ann Thorac Surg*. 2009;88:S23-42.

E3. Shahian DM, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC Jr, et al. The Society of Thoracic Surgeons 2018 adult cardiac surgery risk models: Part 2-statistical methods and results. *Ann Thorac Surg*. 2018; 105:1419-28.

E4. PHC4|Cardiac Surgery in Pennsylvania 2005-2006. [Online]. Accessed March 17, 2023. https://www.phc4.org/reports/cardiac-surgery-in-pennsylvania-2005-2006/

E5. PHC4|Cardiac Surgery in Pennsylvania 2006-2007. [Online]. Accessed March 17, 2023. https://www.phc4.org/reports/cardiac-surgery-in-pennsylvania-2006-2007/

E6. PHC4|Cardiac Surgery in Pennsylvania 2007-2008. [Online]. Accessed March 17, 2023. https://www.phc4.org/reports/cardiac-surgery-in-pennsylvania-2007-2008/

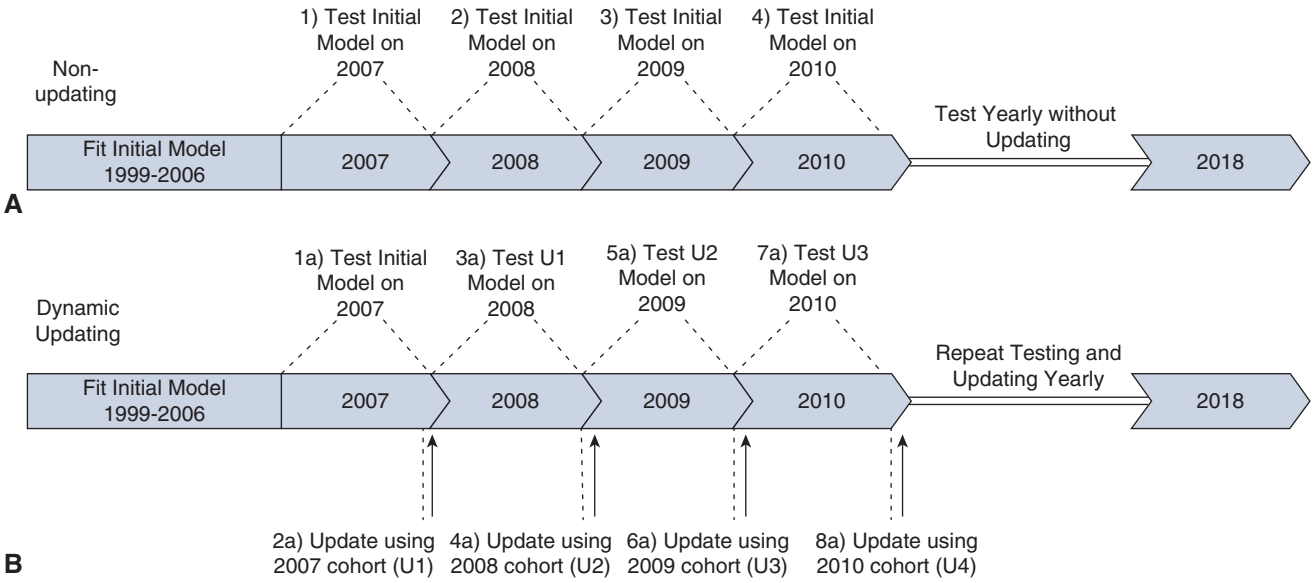


FIGURE E1. Overview of model updating strategies. This figure is a visual representation of model updating compared to not updating a clinical prediction model. Panel A (top) depicts the static approach while panel B (bottom) reflects dynamic updating. The initial model is either the least absolute shrinkage and selection operator (LASSO) or the dynamic logistic state space model (DLSSM) model. In the nonupdating scenario (A), the coefficients from the LASSO model are applied to each subsequent year in the validation set starting in 2007. With dynamic updating (B), the initial model is applied to the 2007 cohort (1a) and then updated using the 2007 data. That updated model (2a) is then tested on the 2008 cohort (3a). The model is then updated a second time (4a) based on the 2008 population and is then tested in the 2009 population (5a). The process of testing followed by updating continues yearly in the validation set.

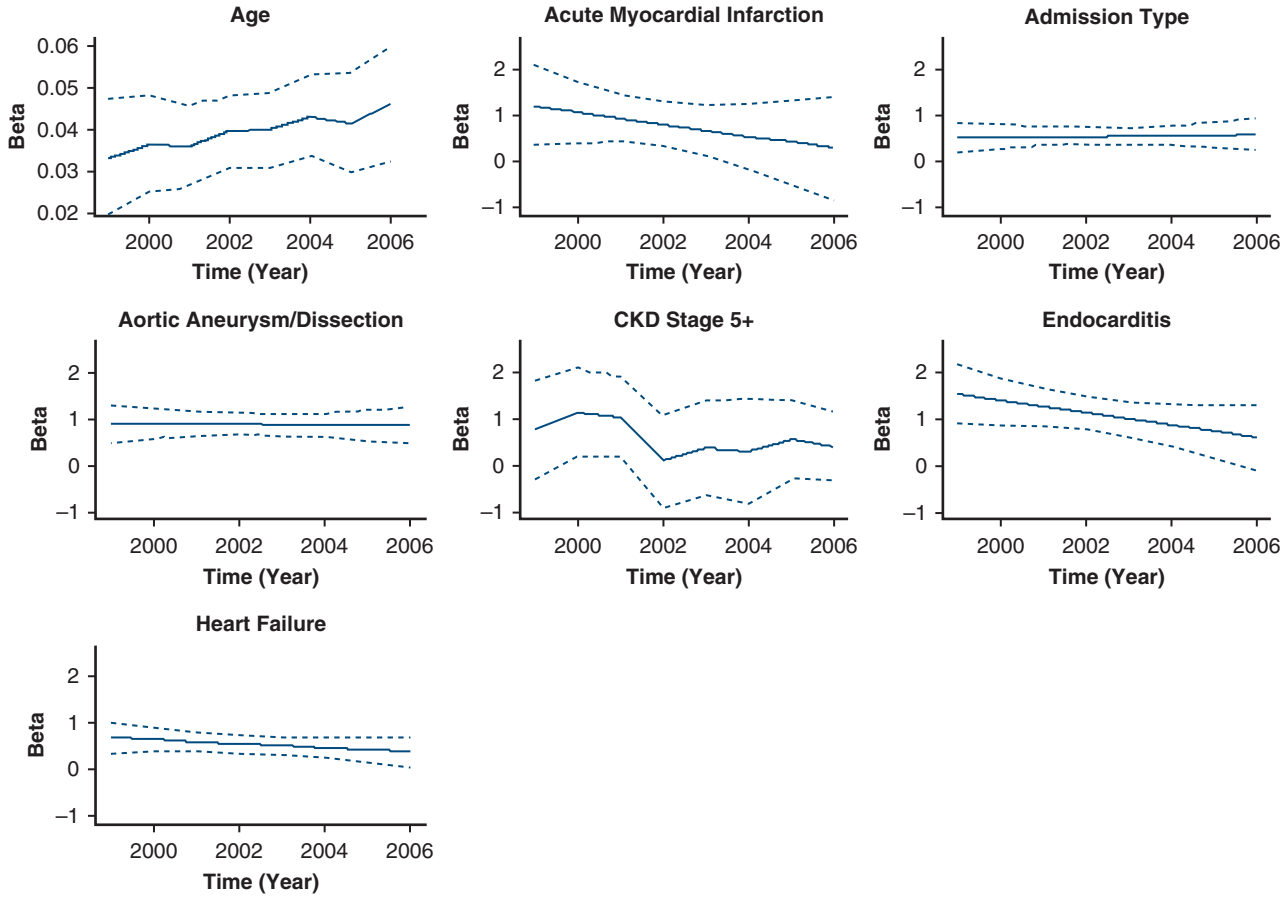


FIGURE E2. Variable selection process of time-varying coefficients for the dynamic logistic state space model (DLSSM). The following plots illustrate the DLSSM variable selection process with corresponding 95% confidence bands (*blue*). The *y*-axis is the beta coefficient for each variable throughout the training period when it is not held constant. Any variable in which the 95% confidence bands excluded the initial point estimate would be considered a meaningful change from the time-invariant model and could be used to inform future changes in the testing set.Note that the *y*-axis varies based on the value of the beta coefficient for age. Due to the limited data and few deaths among those with chronic liver disease, there was not enough information for the model to learn from in the training set. Therefore, we made the assumption that chronic liver disease is a time-invariant coefficient, and the plot is not shown. *CKD*, Chronic kidney disease.

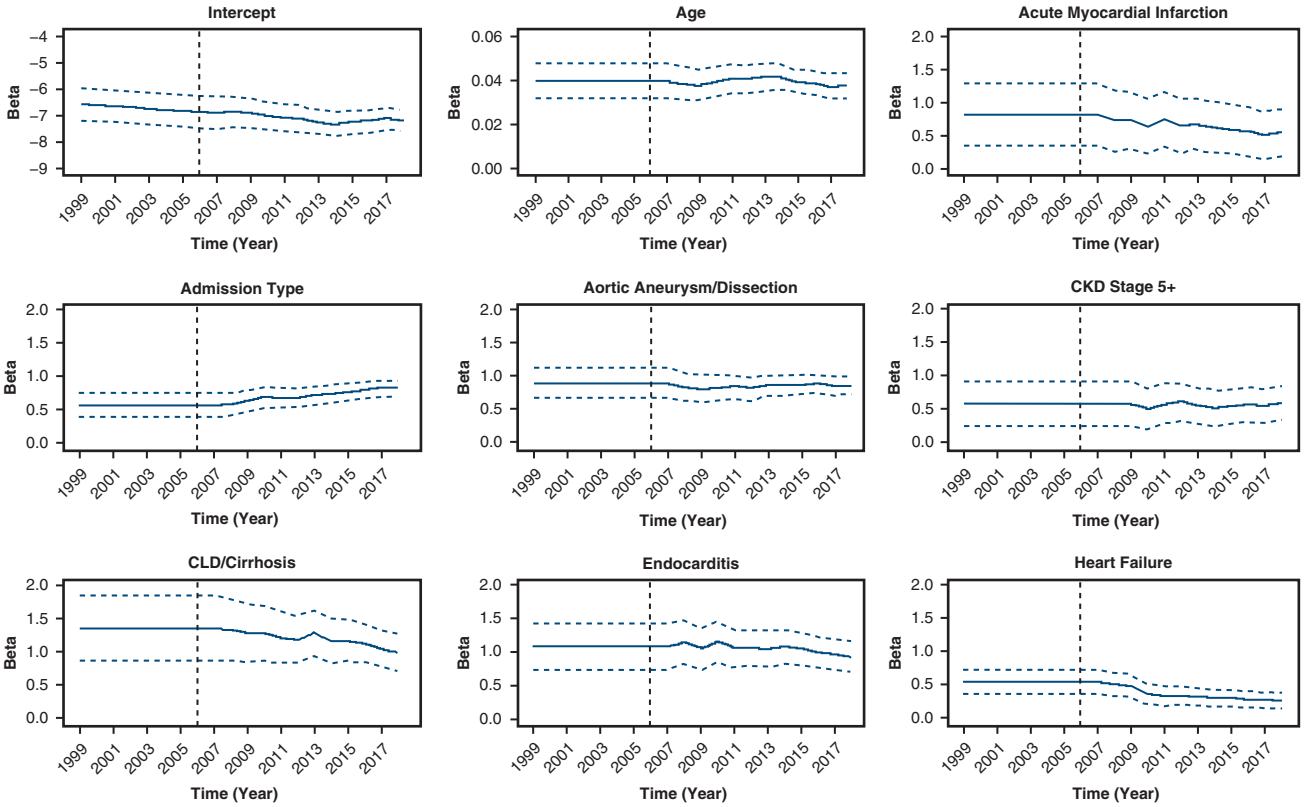


FIGURE E3. Dynamic logistic state space model (DLSSM) final model coefficients. Illustration of DLSSM model coefficients with corresponding 95% confidence bands. The *x*-axis is calendar time, with the *black vertical line* marking the end of the training period (2006) and the beginning of the validation period (2007). The *left side of the vertical line* is the smoothed coefficient in the training set and the *right side* represents the k-step ahead (1 year) prediction in the validation set. In this model, only the intercept is time-varying while the 8 coefficients are held constant (except for random error) within the training set. This figure is significant because it demonstrates which variables would be considered as time-varying and which were in-variant DLSSM model used in the analysis. Note that the *y*-axes vary based on the value of the beta coefficient for the intercept and age plots. *CKD*, Chronic kidney disease; *CLD*, chronic liver disease.

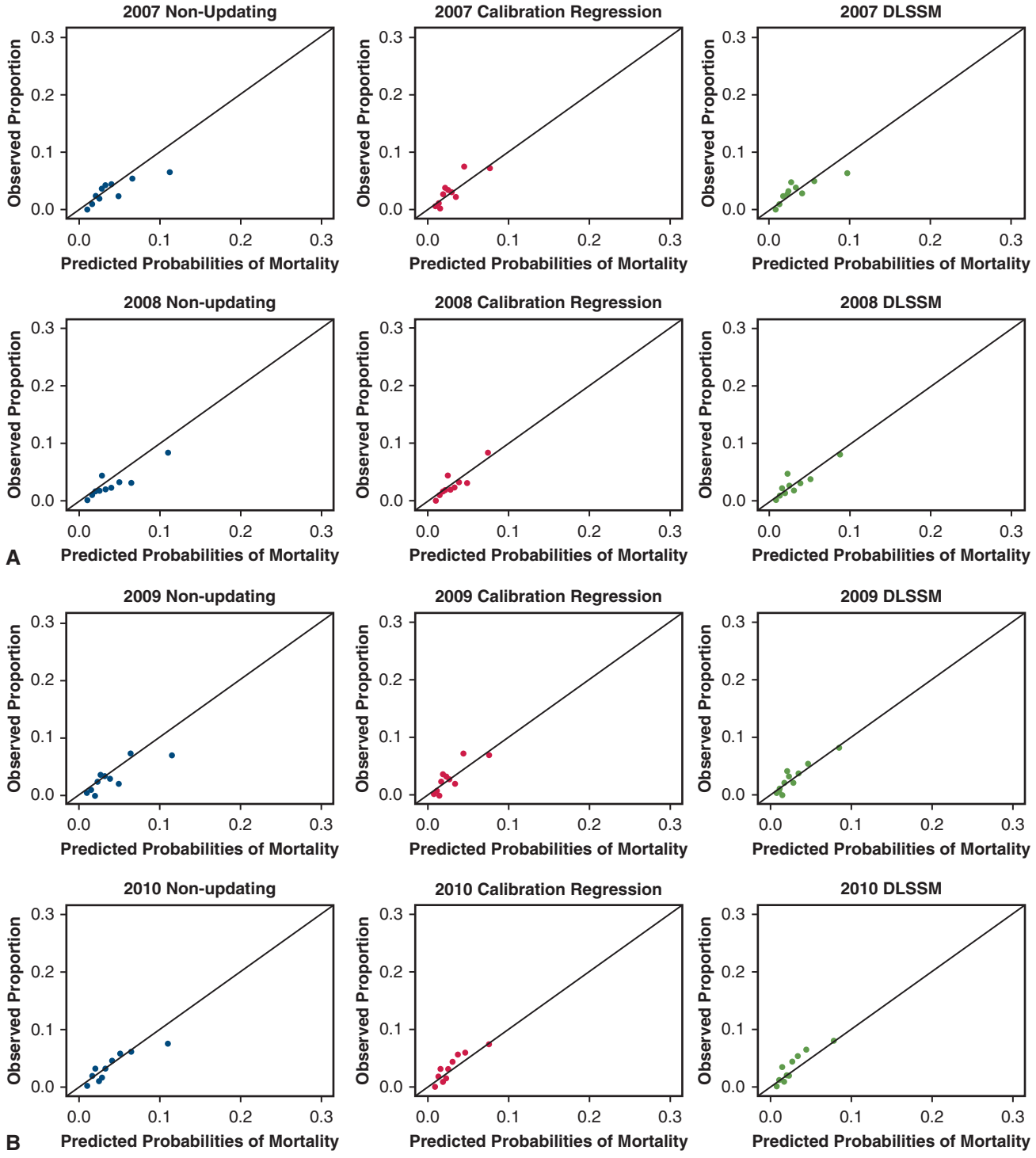


FIGURE E4. Yearly calibration plots by updating strategy (2007-2008). This figure shows the predicted (*x*-axis) by observed (*y*-axis) mortality by risk deciles in the validation sample within each year of the validation set (2007-2018). Because there are 12 years' data, the plots have been divided into 6 total subsets but collectively make up all the yearly calibration plots. A, 2007 to 2008. B, Yearly calibration plots by updating strategy (2009-2010). C, Yearly calibration plots by updating strategy (2011-2012). D, Yearly calibration plots by updating strategy (2013-2014). E, Yearly calibration plots by updating strategy (2015-2016). F, Yearly calibration plots by updating strategy (2017-2018). *DLSSM*, Dynamic logistic state space model.

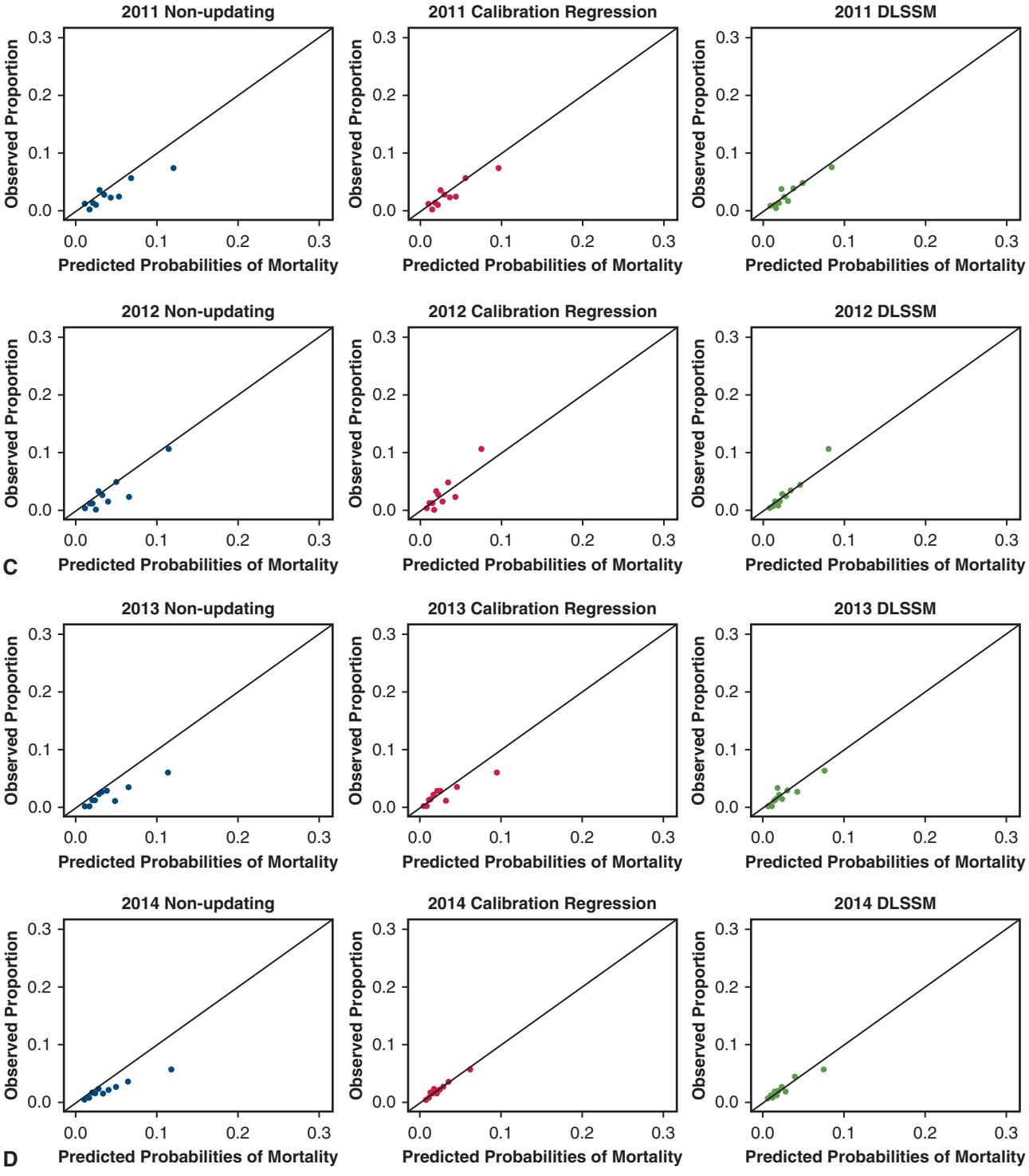
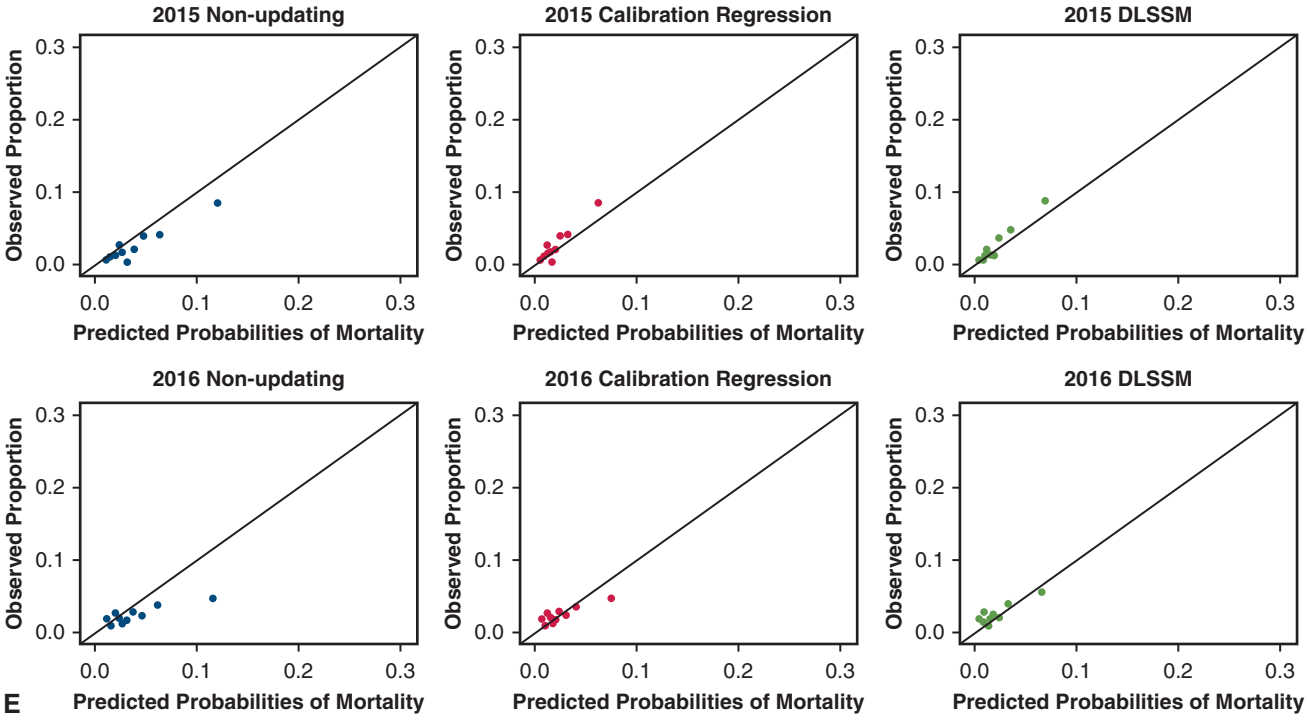


FIGURE E4. Continued.



E

FIGURE E4. Continued.

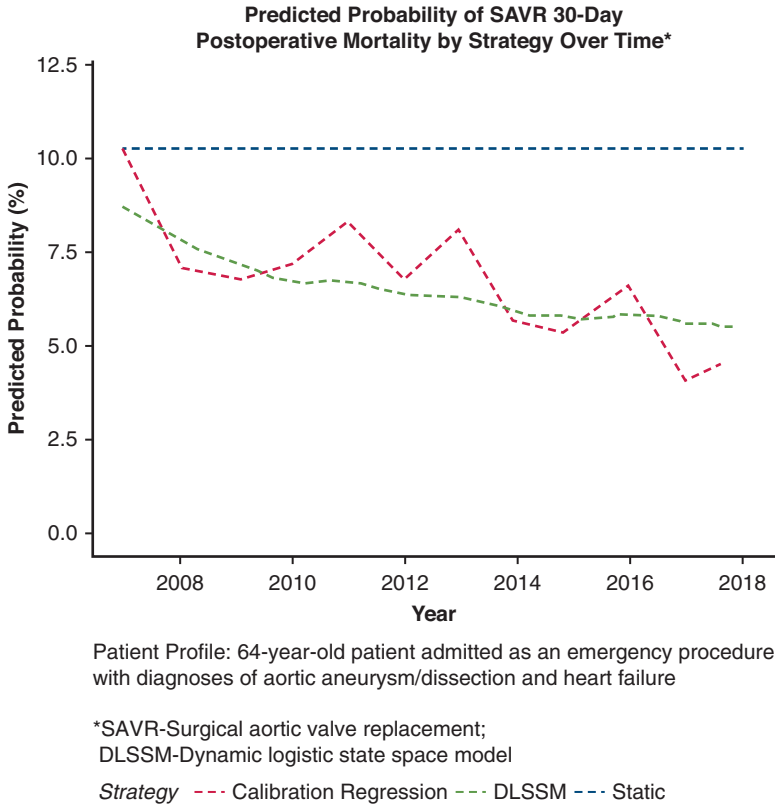


FIGURE E5. Predicted probability of surgical aortic valve replacement (SAVR) 30-day postoperative mortality by strategy over time. The plot shows a comparison of the predicted risk (%) of 30-day postoperative SAVR mortality (*y*-axis) through the validation years of 2007 to 2018 (*x*-axis) by the 3 updating strategies. The *green line* is the static, nonupdating approach. The *blue line* represents the calibration regression strategy, and the *red line* is the dynamic logistic state space model (DLSSM) method. The figure is significant because it demonstrates the temporal trends of predicted risk from each updating strategy. The static approach maintains a fixed risk estimate throughout the study period. In contrast, the calibration regression and DLSSM methods incorporate updated information to adapt their risk predictions over time. *SAVR*, Surgical aortic valve replacement; *DLSSM*, dynamic logistic state space model.

TABLE E1. Subset of univariable analysis: Variables with a protective association with 30-day postoperative mortality

Predictor (n within the 14,070 training sample)	Cross tabulation of mortality with variable (%)	χ^2 value* (*P* value)	Odds ratio (95% CI)
Diabetes (2,327)	62 (2.7)	11.88 (<.001)	0.62 (0.46-0.81)
Depression (452)	6 (1.3)	7.80 (.005)	0.32 (0.12-0.70)
Hypercholesterolemia (3,578)	64 (1.8)	58.67 (<.001)	0.37 (0.28-0.48)
Hypertension (7,308)	198 (2.7)	61.75 (<.001)	0.50 (0.41-0.59)

*With Yates' continuity correction.

TABLE E2. Comparison of model performance across all years in the validation set

		Update year												
Model comparisons	**Metric**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2007-2018**
Calibration														
Nonupdating (static)	H-L	11.827	15.687	18.282	13.086	19.117	23.049	26.748	26.473	18.864	26.557	22.425	31.320	156.490
Calibration regression	H-L	11.827	10.342	16.665	13.875	10.799	17.771	12.443	2.257	12.864	13.132	8.250	25.628	4.491
DLSSM	H-L	9.960	11.638	12.224	18.296	7.734	4.292	8.689	3.630	6.832	16.475	5.059	17.099	7.999
Nonupdating (static)	Intercept*	−0.839	−0.648	−0.641	−0.371	−0.612	0.251	−1.004	−0.852	−0.579	−2.058	−1.074	−0.777	−0.737
Calibration regression	Intercept*	−0.839	0.306	−0.006	0.306	−0.245	1.062	−1.178	0.246	0.291	−1.750	2.422	0.385	−0.273
DLSSM	Intercept*	−0.698	−0.415	−0.189	−0.013	−0.110	0.911	−0.447	−0.257	0.297	−1.425	−0.203	0.215	−0.202
Non-updating (static)	Slope†	0.796	0.906	0.886	0.936	0.927	1.228	0.852	0.932	0.951	0.505	0.858	0.929	0.897
Calibration regression	Slope†	0.796	1.137	0.979	1.056	0.990	1.324	0.694	1.094	1.021	0.531	1.699	1.082	0.933
DLSSM	Slope†	0.796	0.909	0.932	0.937	0.973	1.277	0.901	0.962	1.035	0.589	0.943	1.022	0.935
Discrimination														
Nonupdating (static)	C-Statistic	0.657	0.680	0.705	0.694	0.687	0.738	0.679	0.677	0.703	0.606	0.689	0.673	0.685
Calibration regression	C-Statistic	0.657	0.680	0.705	0.694	0.687	0.738	0.679	0.677	0.703	0.606	0.689	0.673	0.687
DLSSM	C-Statistic	0.656	0.679	0.710	0.698	0.685	0.741	0.685	0.681	0.720	0.617	0.700	0.683	0.696
Accuracy														
Nonupdating (static)	BS	0.0305	0.0268	0.0290	0.0319	0.0280	0.0259	0.0228	0.0213	0.0254	0.0233	0.0208	0.0226	0.0256
Calibration regression	BS	0.0305	0.0265	0.0286	0.0318	0.0277	0.0258	0.0225	0.0207	0.0253	0.0226	0.0201	0.0222	0.0253
DLSSM	BS	0.0303	0.0266	0.0286	0.0319	0.0276	0.0257	0.0222	0.0206	0.0249	0.0225	0.0200	0.0221	0.0252
Nonupdating (static)	MAE	0.067	0.063	0.066	0.069	0.067	0.063	0.059	0.059	0.062	0.059	0.057	0.060	0.063
Calibration regression	MAE	0.067	0.057	0.056	0.060	0.061	0.052	0.048	0.043	0.045	0.047	0.042	0.043	0.052
DLSSM	MAE	0.062	0.056	0.056	0.058	0.056	0.052	0.047	0.044	0.046	0.043	0.040	0.042	0.050

H-L, Hosmer-Lemeshow statistic; *DLSSM*, dynamic-logistic state space model; *BS*, Brier score; *MAE*, mean absolute error. *The intercept measures calibration-in-the-large and refers to the difference between mean expected and mean observed mortality. Values closer to 0 are better with 0 indicating a perfectly calibrated model. †A slope of 1 is a perfectly calibrated model.