Dynamic Prediction Modeling of Postoperative Mortality among Surgical Aortic Valve Replacement Patients in a State-Wide Cohort over a 12-year Period

Jackie Pollack, MSc, Wei Yang, Ph.D., Erin M. Schnellinger, Ph.D., MS, George J. Arnaoutakis, MD, Michael J. Kallan, M.S, Stephen E. Kimmel, MD, MSCE

PII: S2666-2736(23)00197-3
DOI: https://doi.org/10.1016/j.xjon.2023.07.011
Reference: XJON 828

To appear in: JTCVS Open

Received Date: 13 March 2023
Accepted Date: 21 June 2023


This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Copyright © 2023 The Authors. Published by Elsevier Inc. on behalf of The American Association for Thoracic Surgery
Title: Dynamic Prediction Modeling of Postoperative Mortality among Surgical Aortic Valve Replacement Patients in a State-Wide Cohort over a 12-year Period

Authors: Jackie Pollack 1, MSc, Wei Yang 2, Ph.D., Erin M. Schnellinger 3, Ph.D., MS, George J. Arnaoutakis 4, MD, Michael J. Kallan 5, M.S, Stephen E. Kimmel 1 MD, MSCE

Author Affiliation
1. Department of Epidemiology, College of Public Health and Health Professions and College of Medicine, University of Florida, Gainesville, FL 32610
2. Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104
3. Research Department, United Network For Organ Sharing, Richmond, VA 23219
4. Division of Cardiovascular and Thoracic Surgery, University of Texas at Austin Dell Medical School, Austin, Texas
5. Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

Disclosure Statement: The authors declare that they have no conflicts of interest or financial disclosures to report regarding this manuscript.

Funding Sources: This work was funded by the NIH R01HL14129. Ms. Pollack is supported by a Graduate Assistantship in the Department of Epidemiology.

Article word count- 3,481/3,500

Corresponding Author: Stephen E. Kimmel, MD, MSCE (skimmel@ufl.edu). University of Florida, Department of Epidemiology, 2004 Mowry Road, PO Box 100231, Gainesville, Fl 32610

Glossary of Abbreviations
Central Picture Legend: Comparison of 30-Day Postoperative Mortality: Predicted Probabilities and Observed Proportions by Risk Tertiles and Updating Strategies

**Central message (190/200-character limit including spaces)**

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 statement (387/405-character limit including spaces)**

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.

**Abstract (250/250-word limit)**

**Objective:** Clinical prediction models for surgical aortic valve replacement (SAVR) 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 SAVR from January 1, 1999-December 31, 2018, state-wide in Pennsylvania. After chronologically splitting the data into training and validation sets, we compared calibration, discrimination, and accuracy measures amongst a non-updating model to two methods of model updating: calibration regression (CR) and the novel dynamic logistic state space model (DLSSM).

**Results:** The risk of mortality decreased significantly during the validation period (p<0.01) and the non-updating model demonstrated poor calibration and reduced accuracy over time. Both updating models maintained better calibration (Hosmer-Lemeshow [H-L] chi-square statistic) than the non-updating model: non-updating (156.5), calibration regression (4.9), and DLSSM (8.0). Overall accuracy (Brier score) was consistently better across both updating models: DLSSM model (0.0252), calibration regression (0.0253), and non-updating (0.0256).

Discrimination improved with the DLSSM model (AUC=0.696) compared to the non-updating model (AUC=0.685) and calibration regression method (AUC=0.687).

**Conclusion:** 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. As competing therapies have emerged for valve replacement, model updating may guide clinical decision-making.

**Keywords (3-7):** Clinical prediction model, model updating, model recalibration, surgical aortic valve replacement, dynamic logistic state space model

**Introduction**

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 as estimates of mortality range from 1.0-16.4%.\(^4\)

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. While 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 paper 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 for Thoracic Surgeons (STS) models. With decreasing mortality trends for SAVR procedures, evolving care practices, and shifting patient demographics, worsening performance is a natural limitation of static models. 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 last few years, TAVR became the dominant treatment option for AS, even for those with intermediate mortality risk.

Dynamic prediction models (DPMs) are a proposed solution that incorporate underlying changes over time. While 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. 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 two dynamic model updating methods: CR and DLSSM in predicting 30-day postoperative mortality among SAVR patients from the state of Pennsylvania.

We hypothesize that CR and DLSSM will outperform the non-updating 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 non-federal 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. International Classification of Disease Codes (ICD)-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 30 years or older who underwent an isolated SAVR from January 1, 1999-December 31, 2018, in the state of Pennsylvania were included. This period was selected because it: 1) ensured sufficient sample size and study power, 2) incorporated temporal changes and medical advances in the treatment of AS that likely impact one’s estimated probability of survival following the surgery (e.g., TAVR), and 3) provided complete follow-up data at the time of the study. Non-residents were excluded as 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 as they are rare events in this population and affected 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 1999-2006. The validation set contained 30,476 participants between 2007 and 2018.

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 pre-operative events and peri-operative/post-operative complications (arterial embolism and thrombosis, atrial fibrillation/flutter, heart block, pulmonary embolism, stroke, ventricular fibrillation/flutter). Univariable analyses between candidate predictors and the outcome were conducted. At this stage, we found four variables (diabetes, depression, hypertension, and hypercholesterolemia) to have an unexplainable, significant protective association with 30-day mortality (see Supplemental S. Table 1 for sensitivity analysis and further discussion). As 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 (SMD) was calculated for each variable. Values ≥ 0.1 were considered meaningful differences. We compared three approaches: the standard (static) non-updating approach, model updating via CR, and DLSSM. Models were developed in the training set and evaluated for performance in the validation set.
A) Model Development and Updating

A.1) Non-updating 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, S. Model Development). We refer to this model as the “LASSO model”. In this approach, the LASSO model is unchanged in the validation set.

A.2) Logistic Calibration Regression

Logistic CR starts with the LASSO model (i.e., 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 Supplement, S. Figure 1, and S. Methodological Overview for further details).

A.3) Dynamic Logistic State Space Model

For the DLSSM model, we used the DLSSM R package 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 one of the eight 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 eight covariates is time-varying (see Supplement
S.Figure 3). Therefore, the final DLSSM model from the training set included only the time-
varying intercept; the other coefficients remain time-invariant.

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 et al (2021) and the supplemental material (S. Overview of DLSSM).

**B) 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. 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).

Data analyses and graphical outputs were performed using R Version 4.1.2. The study
was reviewed by the University of Florida’s Internal Review Board and received authorization as
non-human 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 1999-2006 (N=14,070, 557 deaths). The validation set included SAVR patients from 2007-2018 (N=30,476, 802 deaths). Each year within the validation set had approximately 2,000-3,000 participants. The annual risk of mortality ranged from approximately 2.0-4.5% and decreased significantly (p<0.001) 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 SAVR patients in the development set was 66.6 years (SD=13.3) and 67.4 years (SD=12.3) in the validation set. Within both sets, SAVR patients were predominately male and undergoing an elective (non-emergency) procedure. In general, patients in the validation set had more comorbidities than those in the training data (SMD>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 eight covariates with similar values between the two models. The model covariates and performance are shown in Table 2.

**Comparison of Updating Strategies**

A) **Calibration**

The static model demonstrated worse calibration, overpredicting the probability of mortality in the validation cohort with a H-L statistic=156.490 (p<0.001) (Figure 3, Table 3). The H-L statistic also predominantly increased each year (Appendix S. Table 2).

The CR and DLSSM models demonstrated better calibration and similar performance in the validation cohort (Table 3, Figure 3). Within each year of updating (S. Table 2), the updating methods demonstrated better calibration (H-L statistic) than the static method, as reflected in the
calibration plots (S. Figure 5). Consistent with these results, the intercept (i.e., calibration-in-the-large) was closer to zero (better calibration) and the slope was closer to one for both updating methods compared with the static method (Table 3). While there was more variability in the year-to-year evaluation of the intercepts and slopes across the three models (S. Table 2), 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 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 (Appendix, S. Figure 4).

B) 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 (S. Table 2). However, when examined across years, CR yielded marginally higher C-statistics (0.687) than non-updating (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 AUCs in most years compared to CR or non-updating models (Supplement, S. Table 2).

C) Overall Accuracy

While differences between the BS are difficult to interpret, the BS was consistently better (lower) in the updating models, both within yearly comparisons and overall (Table 3 and S. Table 2). 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 to CR (MAE=0.052) and non-updating (MAE=0.063). The improvement in MAE progressively widened among the strategies with each successive update (Appendix, S. Table 2).

Discussion

This study evaluated two methods for dynamically updating a CPM in a population of isolated SAVR patients over a period of twelve years and compared these methods with the more typical non-updating approach and with each other. The primary implication of our analyses is that regularly updating a CPM is superior to non-updating. The non-updating model consistently over-predicted 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. While 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 one year, updating through CR does not alter model discrimination. Discrimination is more important when assessing individual risk while improving calibration may have greater impact 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. As 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 one-year post-lung transplant survival, in which DPMs outperformed non-updating. Other studies have also documented dynamically updating models, but they have primarily focused on the estimation of model coefficients rather than prediction accuracy. McCormick et al. (2012) 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. Hickey et al. (2013) 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 Society for Thoracic Surgeons (STS) Adult Cardiac Surgery Database (ACSD) 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 impact 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 prior to development and the process of training, validating, and deploying the model can be substantial. The STS regularly updates the database and applies a year and procedure type-specific correction factor to its institutional reports annually, but the model itself is not updated annually. As a result, the online risk calculator, a decision support tool used by providers throughout the US, is used until a new model is developed. 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. This method only updates the intercept as opposed to CR which updates the intercept and slope.

In 2015, Vassileva et al. compared the 2008 Predicted Risk of Operative Mortality (PROM) 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. The latest 2018 PROM (online calculator version 4.2) showed good calibration and moderate discriminatory ability at the time, but runs the risk of becoming outdated as the model was developed from data 9-12 years ago (between 2011-2014).

Another prominent example in cardiac surgery is the EuroSCORE model which was published in 1999 and tended to overestimate mortality for low-risk SAVR patients and underestimate mortality for high-risk patients in other cardiac surgeries over time. The model was subsequently updated in 2012 (EuroSCORE II) but has not been updated since. Emerging
evidence suggests that EuroSCORE II performance may be deteriorating for some SAVR patients, particularly those ≥ 75 years of age.14

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.33-35 In our previous work with the Lung Allocation Score (LAS), 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 to other updating strategies.21,22 We therefore chose to validate this method here. We were not able to incorporate some common measures used in prominent CPMs of mortality for SAVR patients in our initial models, such as hypertension, hypercholesterolemia, diabetes, or laboratory measurements. This may limit the generalizability of our findings but as 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 to CR (i.e., a more conventional
updating strategy) in a large, state-wide sample of SAVR patients. Second, the PHC4 is a state-
wide agency that includes data from all non-federal 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 LAS.21 Lastly, our outcome measure of 30-day postoperative mortality is robust as
we restricted our sample to Pennsylvania residents and linked patients’ records with death
certificates in Pennsylvania.

Conclusion

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 as 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.

References


16. Report Finds TAVR is Dominant Form of Aortic Valve Replacement, Outcomes Steadily Improving in the United States. *American College of Cardiology*.


Tables
Main Text Table 1 Characteristics of Patient Population Stratified by Development and Validation Cohorts

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Development 1999-2006</th>
<th>Validation 2007-2018</th>
<th>SMD*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Age Years (SD)</td>
<td>66.6 (13.3)</td>
<td>67.4 (12.3)</td>
<td>0.068</td>
</tr>
<tr>
<td>Acute Myocardial Infarction (Primary Diagnosis)</td>
<td>169 (1.2)</td>
<td>187 (0.6)</td>
<td>0.062</td>
</tr>
<tr>
<td>Admission Type (emergency)</td>
<td>4,826 (34.3)</td>
<td>7,822 (25.7)</td>
<td>0.189</td>
</tr>
<tr>
<td>Alzheimer/Dementia</td>
<td>24 (0.2)</td>
<td>172 (0.6)</td>
<td>0.065</td>
</tr>
<tr>
<td>Anemia</td>
<td>2,615 (18.6)</td>
<td>14,791 (48.5)</td>
<td>0.669</td>
</tr>
<tr>
<td>Aortic Aneurysm and/or Dissection</td>
<td>2,150 (15.3)</td>
<td>6,351 (20.8)</td>
<td>0.145</td>
</tr>
<tr>
<td>Aortic Root Surgery (Concomitant)</td>
<td>1,715 (12.2)</td>
<td>5,065 (16.6)</td>
<td>0.126</td>
</tr>
<tr>
<td>Asthma</td>
<td>612 (4.3)</td>
<td>1,736 (5.7)</td>
<td>0.062</td>
</tr>
<tr>
<td>Cardiomyopathy</td>
<td>864 (6.1)</td>
<td>3,043 (10.0)</td>
<td>0.142</td>
</tr>
<tr>
<td>Cachexia</td>
<td>12 (0.1)</td>
<td>47 (0.2)</td>
<td>0.020</td>
</tr>
<tr>
<td>Chronic Pericardial Disease</td>
<td>165 (1.2)</td>
<td>379 (1.2)</td>
<td>0.006</td>
</tr>
<tr>
<td>Chronic Kidney Disease Stage 1-4</td>
<td>254 (1.8)</td>
<td>3,286 (10.8)</td>
<td>0.376</td>
</tr>
<tr>
<td>Chronic Kidney Disease Stage 5+</td>
<td>486 (3.3)</td>
<td>512 (1.7)</td>
<td>0.106</td>
</tr>
<tr>
<td>Chronic Liver Disease</td>
<td>169 (1.2)</td>
<td>678 (2.2)</td>
<td>0.079</td>
</tr>
<tr>
<td>Chronic Obstructive Pulmonary Disease</td>
<td>2,107 (15.0)</td>
<td>4,000 (13.1)</td>
<td>0.053</td>
</tr>
<tr>
<td>Other Chronic Lung Diseases</td>
<td>145 (1.0)</td>
<td>396 (1.3)</td>
<td>0.025</td>
</tr>
<tr>
<td>Coronary Artery Disease</td>
<td>4,017 (28.6)</td>
<td>10,278 (33.7)</td>
<td>0.112</td>
</tr>
<tr>
<td>Endocarditis</td>
<td>465 (3.3)</td>
<td>1,075 (3.5)</td>
<td>0.012</td>
</tr>
<tr>
<td>Excision of Other Lesion/Heart Tissue Same Day</td>
<td>325 (2.3)</td>
<td>1,754 (5.8)</td>
<td>0.176</td>
</tr>
<tr>
<td>Heart Failure</td>
<td>4,834 (34.4)</td>
<td>10,307 (33.8)</td>
<td>0.011</td>
</tr>
<tr>
<td>History of Chronic Steroid Use</td>
<td>13 (0.1)</td>
<td>199 (0.7)</td>
<td>0.092</td>
</tr>
<tr>
<td>Lupus</td>
<td>65 (0.5)</td>
<td>141 (0.5)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Oxygen Dependence Therapy</td>
<td>26 (0.2)</td>
<td>518 (1.7)</td>
<td>0.157</td>
</tr>
<tr>
<td>Obstructive Sleep Apnea</td>
<td>51 (0.4)</td>
<td>2,935 (9.6)</td>
<td>0.435</td>
</tr>
<tr>
<td>Parkinson’s</td>
<td>46 (0.3)</td>
<td>137 (0.4)</td>
<td>0.020</td>
</tr>
<tr>
<td>Peripheral Arterial Disease</td>
<td>2,160 (15.4)</td>
<td>6,477 (21.3)</td>
<td>0.153</td>
</tr>
<tr>
<td>PTCA/Stent</td>
<td>375 (2.7)</td>
<td>1,379 (4.5)</td>
<td>0.100</td>
</tr>
<tr>
<td>Pulmonary Hypertension</td>
<td>405 (2.9)</td>
<td>855 (2.8)</td>
<td>0.004</td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td>124 (0.9)</td>
<td>340 (1.1)</td>
<td>0.024</td>
</tr>
<tr>
<td>Sex (Female)</td>
<td>6,067 (43.1)</td>
<td>11,956 (39.2)</td>
<td>0.079</td>
</tr>
</tbody>
</table>

*SMD-Standardized Mean Difference. Values ≥ 0.1 were considered meaningful differences between the development and validation cohorts and appear in bold font.
Main Text Table 2 Specification of Logistic Regression Models Developed from the Training Data (Years: 1999-2006, N=14,070)

<table>
<thead>
<tr>
<th>Variable</th>
<th>LASSO Derived Models</th>
<th>DLSSM Derived Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\beta) Coefficient**</td>
<td>OR (95% CI)(^*)</td>
</tr>
<tr>
<td>Age</td>
<td>0.04</td>
<td>1.04 (1.03-1.05)</td>
</tr>
<tr>
<td>Acute Myocardial Infarction (Primary diagnosis)</td>
<td>0.82</td>
<td>2.27 (1.37-3.60)</td>
</tr>
<tr>
<td>Admission Type (elective v emergency)</td>
<td>0.57</td>
<td>1.77 (1.47-2.12)</td>
</tr>
<tr>
<td>Aortic Aneurysm and/or Dissection</td>
<td>0.88</td>
<td>2.42 (1.93-3.01)</td>
</tr>
<tr>
<td>Chronic Kidney Disease 5+</td>
<td>0.54</td>
<td>1.72 (1.20-2.40)</td>
</tr>
<tr>
<td>Chronic Liver Disease/Cirrhosis</td>
<td>1.22</td>
<td>3.40 (1.99-5.50)</td>
</tr>
<tr>
<td>Endocarditis</td>
<td>1.06</td>
<td>2.88 (1.97-4.10)</td>
</tr>
<tr>
<td>Heart Failure</td>
<td>0.54</td>
<td>1.71 (1.43-2.05)</td>
</tr>
</tbody>
</table>

Model Performance

<table>
<thead>
<tr>
<th></th>
<th>LASSO Derived Models</th>
<th>DLSSM Derived Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosmer-Lemeshow Statistic (H-L)</td>
<td>9.36***</td>
<td>11.703***</td>
</tr>
<tr>
<td>C-Statistic</td>
<td>0.724</td>
<td>0.750</td>
</tr>
<tr>
<td>Brier Score</td>
<td>0.037</td>
<td>0.034</td>
</tr>
<tr>
<td>Mean Absolute Error</td>
<td>0.074</td>
<td>0.066</td>
</tr>
</tbody>
</table>

The LASSO derived model was used for the non-updating and calibration regression models. The DLSSM model uses the same variables identified by the LASSO and allows the intercept to change over time.

**OR-Odds Ratio, CI-Confidence Interval

**The intercept coefficient for the LASSO model is -6.73. The intercept for the DLSSM model varied over time. For more details refer to supplemental S.Figure 2

***H-L both p-values >0.15
Main Text Table 3 Comparison of Model Performance in the Validation Set

<table>
<thead>
<tr>
<th>Performance Metrics</th>
<th>2007-2018</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-updating</td>
<td>Calibration Regression</td>
</tr>
<tr>
<td><strong>Calibration</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hosmer-Lemeshow</td>
<td>156.490</td>
<td>4.491</td>
</tr>
<tr>
<td>Intercept*</td>
<td>-0.737</td>
<td>-0.273</td>
</tr>
<tr>
<td>Slope*</td>
<td>0.897</td>
<td>0.933</td>
</tr>
<tr>
<td><strong>Discrimination</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-Statistic</td>
<td>0.685</td>
<td>0.687</td>
</tr>
<tr>
<td><strong>Overall Accuracy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brier Score</td>
<td>0.0256</td>
<td>0.0253</td>
</tr>
<tr>
<td>Mean Absolute Error</td>
<td>0.063</td>
<td>0.052</td>
</tr>
</tbody>
</table>

H-L-Hosmer Lemeshow statistic, 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.
Supplemental Table 1 Subset of Univariable Analysis: Variables with a Protective Association with 30-day Postoperative Mortality

<table>
<thead>
<tr>
<th>Predictor</th>
<th>n within the 14,070 training sample</th>
<th>Cross tabulation of mortality with variable (%)</th>
<th>Chi-square value* (p-value)</th>
<th>OR (95% CI)**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>2,327</td>
<td>62 (2.7%)</td>
<td>11.88 (&lt;0.001)</td>
<td>0.62 (0.46-0.81)</td>
</tr>
<tr>
<td>Depression</td>
<td>452</td>
<td>6 (1.3%)</td>
<td>7.80 (0.005)</td>
<td>0.32 (0.12-0.70)</td>
</tr>
<tr>
<td>Hypercholesterolemia</td>
<td>3,578</td>
<td>64 (1.8%)</td>
<td>58.67 (&lt;0.001)</td>
<td>0.37 (0.28-0.48)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>7,308</td>
<td>198 (2.7%)</td>
<td>61.75 (&lt;0.001)</td>
<td>0.50 (0.41-0.59)</td>
</tr>
</tbody>
</table>

*With Yates’ continuity correction

**OR=odds ratio, CI=confidence interval
### Supplemental Table 2 Comparison of Model Performance Across All Years in the Validation Set

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-updating (static)</td>
<td>Intercept*</td>
<td>-0.839</td>
<td>-0.648</td>
<td>-0.641</td>
<td>-0.371</td>
<td>-0.612</td>
<td>-1.004</td>
<td>-0.852</td>
<td>-0.579</td>
<td>-2.058</td>
<td>-1.074</td>
<td>-0.777</td>
<td>-0.737</td>
<td></td>
</tr>
<tr>
<td>Calibration Regression</td>
<td>Intercept*</td>
<td>-0.839</td>
<td>0.306</td>
<td>-0.006</td>
<td>0.306</td>
<td>-0.245</td>
<td>1.062</td>
<td>-1.178</td>
<td>0.246</td>
<td>0.291</td>
<td>-1.750</td>
<td>2.422</td>
<td>0.385</td>
<td>-0.273</td>
</tr>
<tr>
<td>DLSSM</td>
<td>Intercept*</td>
<td>-0.698</td>
<td>-0.415</td>
<td>-0.189</td>
<td>-0.013</td>
<td>-0.110</td>
<td>0.911</td>
<td>-0.447</td>
<td>-0.257</td>
<td>0.297</td>
<td>-1.425</td>
<td>-0.203</td>
<td>0.215</td>
<td>-0.202</td>
</tr>
<tr>
<td>Non-updating (static)</td>
<td>Slope*</td>
<td>0.796</td>
<td>0.906</td>
<td>0.886</td>
<td>0.936</td>
<td>0.927</td>
<td>1.228</td>
<td>0.852</td>
<td>0.932</td>
<td>0.951</td>
<td>0.505</td>
<td>0.858</td>
<td>0.929</td>
<td>0.897</td>
</tr>
<tr>
<td>Calibration Regression</td>
<td>Slope*</td>
<td>0.796</td>
<td>1.137</td>
<td>0.979</td>
<td>1.056</td>
<td>0.990</td>
<td>1.324</td>
<td>0.694</td>
<td>1.094</td>
<td>1.021</td>
<td>0.531</td>
<td>1.699</td>
<td>1.082</td>
<td>0.933</td>
</tr>
<tr>
<td>DLSSM</td>
<td>Slope*</td>
<td>0.796</td>
<td>0.909</td>
<td>0.932</td>
<td>0.937</td>
<td>0.973</td>
<td>1.277</td>
<td>0.901</td>
<td>0.962</td>
<td>1.035</td>
<td>0.589</td>
<td>0.943</td>
<td>1.022</td>
<td>0.935</td>
</tr>
<tr>
<td>Discrimination</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-updating (static)</td>
<td>C-Statistic</td>
<td>0.657</td>
<td>0.680</td>
<td>0.705</td>
<td>0.694</td>
<td>0.687</td>
<td>0.738</td>
<td>0.679</td>
<td>0.677</td>
<td>0.703</td>
<td>0.606</td>
<td>0.689</td>
<td>0.673</td>
<td>0.685</td>
</tr>
<tr>
<td>Calibration Regression</td>
<td>C-Statistic</td>
<td>0.657</td>
<td>0.680</td>
<td>0.705</td>
<td>0.694</td>
<td>0.687</td>
<td>0.738</td>
<td>0.679</td>
<td>0.677</td>
<td>0.703</td>
<td>0.606</td>
<td>0.689</td>
<td>0.673</td>
<td>0.687</td>
</tr>
<tr>
<td>DLSSM</td>
<td>C-Statistic</td>
<td>0.656</td>
<td>0.679</td>
<td>0.710</td>
<td>0.698</td>
<td>0.685</td>
<td>0.741</td>
<td>0.685</td>
<td>0.681</td>
<td>0.720</td>
<td>0.617</td>
<td>0.700</td>
<td>0.683</td>
<td>0.696</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-updating (static)</td>
<td>BS</td>
<td>0.0305</td>
<td>0.0268</td>
<td>0.0290</td>
<td>0.0319</td>
<td>0.0280</td>
<td>0.0259</td>
<td>0.0228</td>
<td>0.0213</td>
<td>0.0254</td>
<td>0.0233</td>
<td>0.0208</td>
<td>0.0226</td>
<td>0.0256</td>
</tr>
<tr>
<td>Calibration Regression</td>
<td>BS</td>
<td>0.0305</td>
<td>0.0265</td>
<td>0.0286</td>
<td>0.0318</td>
<td>0.0277</td>
<td>0.0258</td>
<td>0.0225</td>
<td>0.0207</td>
<td>0.0253</td>
<td>0.0226</td>
<td>0.0201</td>
<td>0.0222</td>
<td>0.0253</td>
</tr>
<tr>
<td>DLSSM</td>
<td>BS</td>
<td>0.0303</td>
<td>0.0266</td>
<td>0.0286</td>
<td>0.0319</td>
<td>0.0276</td>
<td>0.0257</td>
<td>0.0222</td>
<td>0.0206</td>
<td>0.0249</td>
<td>0.0225</td>
<td>0.0200</td>
<td>0.0221</td>
<td>0.0252</td>
</tr>
<tr>
<td>Non-updating (static)</td>
<td>MAE</td>
<td>0.067</td>
<td>0.063</td>
<td>0.066</td>
<td>0.069</td>
<td>0.067</td>
<td>0.063</td>
<td>0.059</td>
<td>0.059</td>
<td>0.062</td>
<td>0.059</td>
<td>0.057</td>
<td>0.060</td>
<td>0.063</td>
</tr>
<tr>
<td>Calibration Regression</td>
<td>MAE</td>
<td>0.067</td>
<td>0.057</td>
<td>0.056</td>
<td>0.060</td>
<td>0.061</td>
<td>0.052</td>
<td>0.048</td>
<td>0.043</td>
<td>0.045</td>
<td>0.047</td>
<td>0.042</td>
<td>0.043</td>
<td>0.052</td>
</tr>
<tr>
<td>DLSSM</td>
<td>MAE</td>
<td>0.062</td>
<td>0.056</td>
<td>0.056</td>
<td>0.058</td>
<td>0.056</td>
<td>0.052</td>
<td>0.047</td>
<td>0.044</td>
<td>0.046</td>
<td>0.043</td>
<td>0.040</td>
<td>0.042</td>
<td>0.050</td>
</tr>
</tbody>
</table>

- H-L: Hosmer Lemeshow statistic, 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.
Supplemental Material Text

S. Data Dictionary

A detailed data dictionary of PHC4 data is available online\(^1\) with the hyperlink below. \(^1\)


S. Sensitivity Analysis

Univariable analyses between candidate predictors and our primary outcome were conducted. At this stage, we found four 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 under-coded/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 four conditions may have been missed by being listed as a complication of an existing condition. For example, if an individual had an ICD-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 impact of retaining these variables and fit a model using 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 to 0.72 from our baseline model without them. As 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 while 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.\textsuperscript{2,3}

Finally, these four 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. While 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).\textsuperscript{4-6}

S. Model Development

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.
Two lambda values corresponding to the minimum and one standard error above the minimum
were considered, where the latter provides a more parsimonious model. We compared the
performance of the two 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 (i.e., C-statistic). The C-statistics were similar for the two models (0.74 for the minimum value of lambda and 0.72 for one standard error above), so the more parsimonious model was chosen to test the updating strategies in the validation cohort (Table 2, LASSO-Derived model).

S. 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 two 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:

\[
lp = \alpha_{\text{new}} + \alpha_{\text{original}} + (\beta_{\text{original}} \times x_{i\ldots\text{new}}).
\]

The updated slope (β) is the result of multiplying the slope from the new model by the slope of the original model (β_\text{new} = β_{\text{original}} \times β_{\text{new model}}). Subsequent updates follow the same procedure, using the linear predictor from the most recently updated model.

S Overview of DLSSM

In contrast to CR, 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 STS database, is that DLSSM may improve both discrimination and calibration while 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.

Supplemental Material References

1. PHC4.org. Inpatient Discharge Data File and Supporting Documentation File Layouts (UB04),
   1990- Present. 1990; https://www.phc4.org/services/datarequests/docs/specialrequests1990-
**Figure Legends**

Figure 1: Derivation of Study Sample. Significance-This figure maps the inclusion and exclusion criteria for the population used for this analysis and also shows how the data was split into the development and validation sets.

Figure 2: Risk of 30-day Postoperative Mortality after SAVR in Patient Population. Significance- This figure plots the risk of mortality among surgical aortic valve replacement (SAVR) patients over time throughout the study period.

Figure 3: Calibration Plots of Updating Strategies. Significance- This figure shows the predicted (x-axis) by observed (y-axis) mortality by risk deciles in the validation sample. H-L-Hosmer-Lemeshow Chi-Square statistic with the corresponding p-value.

DLSSM-Dynamic logistic state space model

Supplemental Figure 1: Overview of Model Updating Strategies. Significance- This figure is a visual representation of model updating compared to not updating a clinical prediction mode. The “initial model” is either the Least Absolute Shrinkage and Selection Operator (LASSO) or the dynamic logistic state space model (DLSSM) model. In the non-updating scenario (top), the coefficients from the LASSO model are applied to each subsequent year in the validation set starting in 2007. With dynamic updating, 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. The process of testing followed by updating continues yearly in the validation set.

Supplemental Figure 2: 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 gray 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 eight coefficients are held constant (except for random error) within the training set. This figure is significant as it demonstrates which variables would be considered as time-varying and which ones were invariant 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

Supplemental Figure 3: Variable Selection Process of Time-varying coefficients for DLSSM.

The following plots illustrate the DLSSM variable selection process with corresponding 95% confidence bands (gray). 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.

**CKD=Chronic kidney disease

***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.

Supplemental Figure 4: Predicted Probability of 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 2007-2018 (x-axis) by the three updating strategies. The yellow line is the static, non-updating approach. The light blue line represents the calibration regression strategy, and the dark blue line is the DLSSM method. The figure is significant as 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.

Supplemental Figure 5.1: 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) DLSSM-Dynamic Logistic State Space Model. As there are 12 years of data, the plots have been divided into 6 total subsets but collectively make up all the yearly calibration plots. Supplemental Figure 5.1 contains 2007-2008.
• Supplemental Figure 5.2: Yearly Calibration Plots by Updating Strategy (2009-2010).
  Continuation of Supplemental Figure 5.1

• Supplemental Figure 5.3: Yearly Calibration Plots by Updating Strategy (2011-2012)

• Supplemental Figure 5.4: Yearly Calibration Plots by Updating Strategy (2013-2014)

• Supplemental Figure 5.5: Yearly Calibration Plots by Updating Strategy (2015-2016)

• Supplemental Figure 5.6: Yearly Calibration Plots by Updating Strategy (2017-2018)
Risk SAVR Mortality Over Study Period

Risk of 30-day Postoperative Mortality (%) vs Year
Figure 3 Calibration Plots of Updating Strategies

- **Non-updating (2007-2010)**: H-L p-value <0.001
- **Calibration Regression (2007-2018)**: H-L p-value 0.810
- **DLSSM (2007-2018)**: H-L p-value 0.434

DLSSM - Dynamic logistic state space model
H-L - Hosmer-Lemeshow Chi-Square statistic with corresponding p-value
Predictive Performance of Mortality Models in SAVR Patients

- Static
- Calibration Regression
- DLSSM

Risk Tertile

<table>
<thead>
<tr>
<th>Value (%)</th>
<th>Low</th>
<th>Intermediate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Predicted Probabilities</td>
<td>2.0</td>
<td>3.5</td>
<td>4.5</td>
</tr>
<tr>
<td>Observed Proportion</td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

*SAVR: Surgical Aortic Valve Replacement,
DLSSM: Dynamic Logistic State Space Model
Predicted Probability of SAVR 30-Day Postoperative Mortality by Strategy Over Time*

Patient Profile: 64-year-old patient admitted as an emergency procedure with diagnoses of aortic aneurysm/dissection and heart failure

*SAVR-Surgical aortic valve replacement
DLSSM-Dynamic logistic state space model