Improved Prediction by Dynamic Modeling
An Exploratory Study in the Adult Cardiac Surgery Database of the Netherlands Association for Cardio-Thoracic Surgery

Sabrina Siregar, MD, PhD*; Daan Nieboer, MSc*; Yvonne Vergouwe, PhD; Michel I.M. Versteegh, MD; Luc Noyez, MD, PhD; Alexander B.A. Vonk, MD, PhD; Ewout W. Steyerberg, PhD; Johanna J.M. Takkenberg, MD, PhD

Background—The predictive performance of static risk prediction models such as EuroSCORE deteriorates over time. We aimed to explore different methods for continuous updating of EuroSCORE (dynamic modeling) to improve risk prediction.

Methods and Results—Data on adult cardiac surgery from 2007 to 2012 (n=95,240) were extracted from the Netherlands Association for Cardio-Thoracic Surgery database. The logistic EuroSCORE predicting in-hospital death was updated using 6 methods: recalibrating the intercept of the logistic regression model; recalibrating the intercept and joint effects of the prognostic factors; re-estimating all prognostic factor effects, re-estimating all prognostic factor effects, and applying shrinkage of the estimates; applying a test procedure to select either of these; and a Bayesian learning strategy. Models were updated with 1 or 3 years of data, in all cardiac surgery or within operation subgroups. Performance was tested in the subsequent year according to discrimination (area under the receiver operating curve, area under the curve) and calibration (calibration slope and calibration-in-the-large). Compared with the original EuroSCORE, all updating methods resulted in improved calibration-in-the-large (range ~0.17 to 0.04 versus −1.13 to −0.97, ideally 0.0). Calibration slope (range 0.92–1.15) and discrimination (area under the curve range 0.83–0.87) were similar across methods. In small subgroups, such as aortic valve replacement and aortic valve replacement+coronary artery bypass grafting, extensive updating using 1 year of data led to poorer performance than using the original EuroSCORE. The choice of updating method had little effect on benchmarking results of all cardiac surgery.

Conclusions—Several methods for dynamic modeling may result in good discrimination and superior calibration compared with the original EuroSCORE. For large populations, all methods are appropriate. For smaller subgroups, it is recommended to use data from multiple years or a Bayesian approach. (Circ Cardiovasc Qual Outcomes. 2016;9:171-181. DOI: 10.1161/CIRCOUTCOMES.114.001645.)

Key Words: database ■ epidemiology ■ Euroscore ■ outcome ■ risk prediction ■ statistics ■ surgery

Clinical risk prediction models are widely applied in cardiac surgery. They are used for patient selection and outcomes evaluation within and between centers, and they are indispensable for quality improvement programs. In this context, the predictive performance of prediction models has major consequences. Poor predictions may lead to inadequate risk estimates for patient-level decision making and a misleading perception of the quality of care.

The most commonly used prediction model for mortality after cardiac surgery in Europe is the EuroSCORE. A systematic review published in 2012 showed that the model overestimated mortality because it was published in 1999. In a recent study using the Society for Cardio-Thoracic Surgery database, a drift of the calibration (progressive overestimation of mortality over time) was seen between 2001 and 2011. This illustrates that the predictive performance of prediction models may deteriorate over time. The underlying mechanisms of calibration drift include a change in the patient risk profile and improved outcomes, which both have been observed in large national database studies. In other words, as the patient population ages, the indication for cardiac surgery broadens and cardiac surgical care improves, the relation between prognostic factors and outcomes may gradually change too.

Received December 29, 2014; accepted January 22, 2016. From the Department of Cardio-Thoracic Surgery, Leiden University Medical Center, Leiden, The Netherlands (S.S., M.I.M.V.); Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands (D.N., Y.V., E.W.S.); Department of Cardio-Thoracic Surgery, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands (L.N.); Department of Cardio-Thoracic Surgery, VU Medical Center, Amsterdam, The Netherlands (A.B.A.V.); and Department of Cardio-Thoracic Surgery, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands (J.J.M.T.).

*Dr Siregar and D. Nieboer contributed equally to this work.


Correspondence to Sabrina Siregar, Department of Cardio-Thoracic Surgery, Leiden University Medical Center, post zone K-6, PO box 9600, 2300 RC Leiden, The Netherlands. E-mail s.siregar@lumc.nl

© 2016 American Heart Association, Inc.

Circ Cardiovasc Qual Outcomes is available at http://circoutcomes.ahajournals.org

DOI: 10.1161/CIRCOUTCOMES.114.001645

171
Risk prediction models hence need to be periodically updated, also referred to as dynamic modeling. Various methods can be applied for model updating, which have not been explored and applied in the context of periodical updating of cardiac surgery risk models. The aim of this study was to explore different methods for model updating in the setting of adult cardiac surgery and study the consequences on benchmarking, using a large high-quality national database.

Methods

Data
Data were extracted from the Netherlands Association for Cardio-Thoracic Surgery (Nederlandse Vereniging voor Thoraxchirurgie) database. This national database includes all cardiac surgery performed since 2007. Completeness is exceptionally high, with all 16 centers participating, a low number of missing values (0.01%), and >99% complete cases. Information was used on adult cardiac surgery performed from January 1, 2007, to December 31, 2012 (n=95,240). This comprised patient demographics, in-hospital mortality, and EuroSCORE prognostic factors for mortality: age, female, chronic pulmonary disease, extracardiac arteriopathy, neurological dysfunction, previous cardiac surgery, serum creatinine >200 μmol/L, active endocarditis, critical preoperative state, unstable angina on intravenous nitrates, reduced left ventricular ejection fraction (30% to 50% and <30%), recent (<90 days) myocardial infarction, pulmonary systolic pressure >60 mm Hg, emergency surgery, other than isolated coronary surgery, thoracic aorta surgery, and surgery for postinfarct septal rupture. The data provided for this study were fully anonymized; it was not possible to identify the individuals from the information provided. Therefore, approval from the ethics committee was not required and not obtained.

Updating Methods
We compared 6 different model updating methods for the EuroSCORE using data from either 1 or 3 consecutive years from the the Netherlands Association for Cardio-Thoracic Surgery data set. Updating methods included simple recalibration, that is, re-estimating the model intercept or re-estimating the model intercept and calibration slope, re-estimating individual prognostic factor effects, re-estimating all prognostic factor effects and applying shrinkage, a closed test procedure, and Bayesian learning, as detailed later (Figure 1) and in the Data Supplement. The updated models were validated on data from the subsequent year.

All analyses were first performed in all cardiac surgery patients and then repeated within the following operation subgroups: isolated coronary artery bypass grafting (CABG), aortic valve replacement (AVR), AVR with CABG and all other operations (Other). The prognostic factors that were not available in all operation subgroups were kept fixed during the updating of the EuroSCORE (eg, the prognostic factor surgery on thoracic aorta in the subgroup isolated CABG). In addition, an alternative approach to modeling subgroups was applied by including operation subgroup as a variable (aforementioned groups as categories) in the EuroSCORE model.

Intercept Recalibration (Re-Estimating the Model Intercept)
The simplest form of recalibration of a prediction model is to re-estimate only the model intercept. By using a recalibration factor, this approach ensures that the estimates of the prediction model are on average correct. The recalibration factor is estimated by fitting a logistic regression model with only an intercept and the linear predictor of the logistic Euroscore as an offset variable (prognostic factor in a logistic regression model where the coefficient of the prognostic factor is fixed at unity).

Logistic Recalibration (Re-Estimating the Model Intercept and Calibration Slope)
In the second approach, logistic recalibration, the model intercept and prognostic factor effects are updated by one common factor for all coefficients. This is done by fitting a logistic regression model with the linear predictor of the logistic Euroscore as the only covariate.

Model Revision (Re-Estimating Individual Prognostic Factor Effects)
The third approach for updating is model revision: all individual prognostic factor effects are updated. Model revision thus requires the re-estimation of a larger number of coefficients. In small samples, the estimates of the prognostic factor effects can become unstable. To increase sample sizes, we combined data from multiple years. Model revision was thus performed using 1 year and using 3 consecutive years.

Model Revision With Shrinkage (Re-Estimating Individual Prognostic Factor Effects)
The fourth approach is model revision with shrinkage: here again, all individual prognostic factor effects are updated. After re-estimating the effects, shrinkage toward the recalibrated coefficients is applied to stabilize the prognostic factor estimates. Using these shrinkage techniques, the prognostic factor effects of the updated model will be somewhere between those of model revision and logistic recalibration.

Closed Test Procedure
Rather than choosing one of the aforementioned approaches, we attempted to identify the most suitable method using a statistical test. In this testing procedure, the original EuroSCORE model and the models resulting from intercept recalibration, logistic recalibration, and model revision are tested against each other on model fit (measured using the difference in the −2 log-likelihood). The procedure is called a closed testing procedure because multiple tests were performed while maintaining a prespecified statistical significance level alpha (set to 0.05). Again, to increase sample sizes, the closed test was applied using patient data from 1 year and from 3 consecutive years.

Bayesian Model Updating
Updating of prediction models can be considered as a Bayesian process. The background knowledge available from the EuroSCORE—specified as the prior distribution—is combined with new patient information—specified as the likelihood—to obtain updated estimates of the prognostic factor effects, also called the posterior distribution.
The prior distribution of the prognostic factor effects were independent normal distributions with the prognostic factor effects from the original EuroSCORE model as means. The standard deviation of the normal distributions was set equal to such that the probability was 95% that the odds ratio of each prognostic factor was between 1/4 and 4 times the odds ratio estimated by the EuroSCORE model.16,17

Comparison of Updated Models

Prognostic Factor Effects

The estimates of the prognostic factor effects in the updated EuroSCORE models were assessed and studied across the years. We expected some sudden changes in the estimated prognostic factor effects in consecutive years, which reflect random variation (unstable estimates) rather than real changes in the strength of the prognostic factors.

Validation of Updated Models

The updated models were validated in patient data from the subsequent year by assessing the calibration and discrimination.11 Calibration refers to the ability of a model to reliably estimate the risk of mortality and was assessed using calibration-in-the-large and calibration slope. Calibration-in-the-large refers to the difference between the mean expected and mean observed mortality and is exactly 0 in a perfectly calibrated model. The calibration slope is obtained by fitting a logistic regression model with the linear predictor of a model as a single covariate. The slope is 1 in a perfectly calibrated model. A calibration slope smaller than 1 indicates that predicted risks were too extreme in the sense of overestimating for patients at high risk, while underestimating for patients at low risk and is indicative of overfitting of the model. Discrimination refers to the ability of a model to differentiate between survivors and nonsurvivors and was assessed using the area under the curve (AUC) of the receiver operating curve. The AUC ranges from 0.5 (no ability to discriminate) to 1.0 (perfect ability to discriminate).

Benchmarking

The EuroSCORE can be used to correct for differences in case-mix when comparing mortality rates between hospitals.18 It is expected that an improved updated EuroSCORE model enables a better comparison between hospitals. We used the original EuroSCORE model, the updated EuroSCORE model after model revision, the EuroSCORE model updated using the closed test procedure, and the Bayesian updated EuroSCORE model to compare risk-adjusted mortality rates between hospitals. We compared the hospital-specific risk-adjusted estimates by creating a plot and assessing the outliers.19,20 A hospital was identified as an outlier if its random intercept did not include 0 in the 95% prediction interval. As recalibration does not influence the comparison between hospitals, updating methods based on recalibration were not considered in this analysis.

All analyses were performed in R Statistical Software version 3.1.1. The Bayesian approach was implemented using the function MCMClogit from the R-package MCMCpack. The R code can be found in the Data Supplement.

Table. Number of Operations and Mortality per Year in Each Operation Subgroup

<table>
<thead>
<tr>
<th>Year</th>
<th>CABG</th>
<th>AVR</th>
<th>AVR and CABG</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>8471 (1.52%)</td>
<td>1597 (2.44%)</td>
<td>1145 (4.80%)</td>
<td>4390 (6.36%)</td>
<td>15603</td>
</tr>
<tr>
<td>2008</td>
<td>8519 (1.51%)</td>
<td>1669 (2.34%)</td>
<td>1170 (5.38%)</td>
<td>4388 (6.29%)</td>
<td>15818</td>
</tr>
<tr>
<td>2009</td>
<td>8448 (1.30%)</td>
<td>1701 (2.41%)</td>
<td>1196 (3.01%)</td>
<td>4483 (5.53%)</td>
<td>15828</td>
</tr>
<tr>
<td>2010</td>
<td>8368 (1.28%)</td>
<td>1893 (2.11%)</td>
<td>1201 (3.08%)</td>
<td>4632 (6.00%)</td>
<td>16094</td>
</tr>
<tr>
<td>2011</td>
<td>8347 (1.41%)</td>
<td>1905 (1.47%)</td>
<td>1203 (4.07%)</td>
<td>4651 (6.08%)</td>
<td>16106</td>
</tr>
<tr>
<td>2012</td>
<td>8046 (1.14%)</td>
<td>1819 (1.65%)</td>
<td>1203 (3.33%)</td>
<td>4720 (6.48%)</td>
<td>15788</td>
</tr>
</tbody>
</table>

AVR indicates aortic valve replacement; and CABG, coronary artery bypass grafting.

Results

Patient Characteristics

The total number of interventions increased over the years 2007 to 2012. The number of all types of operations increased, except for the number of isolated CABG procedures (Table). The mean age increased from 69 in 2007 to 70 in 2012, and the median EuroSCORE increased from 3.98% to 4.17% in 2012 (P=0.02). Overall, the mortality rate declined over the years (Table). Mortality rates were lowest in the operation subgroup isolated CABG and highest in the operation subgroup other, which included all interventions that do not belong to the subgroups isolated CABG, isolated AVR, or AVR and CABG.

EuroSCORE Updating for All Cardiac Surgery

Prognostic Factor Effects

We first used all cardiac surgery procedures for model updating. When the updating process was based on 1 year of data, the estimates of prognostic factor effects appeared to be unstable (Figure 2). This was most evident for the model revision method. However, the closed test procedure also resulted in unstable estimates. Model updating using data from 3 consecutive years led to smoother changes in the estimated prognostic factor effects for both model revision and the closed test procedure. Less extensive model updating such as the logistic recalibration method led to fairly stable prognostic factor estimates with 1 year of data. Bayesian learning also showed smooth changes of the prognostic factor effects. The prognostic factor effects of neurological dysfunction, previous cardiac surgery, other than isolated CABG, surgery on thoracic aorta, and postinfarct septal rupture were consistently different from the estimates in the EuroSCORE model. The incidence of postinfarct septal rupture was low, and thus the effect could not be reliably estimated. Using smaller prior variances led to smoother changes in the prognostic factor effects, whereas larger variances led to faster adjustments of the prognostic factor effects. Orthogonalizing the predictor effects led to similar results as using the Bayesian approach without orthogonalizing the predictors (results not shown).

Performance of Updated Models

All updating methods showed an improved calibration-in-the-large compared with the EuroSCORE (Figure 3), ranging from −0.17 to 0.04. The calibration slope of the updated models and the original model ranged from −0.92 to 1.15, and the AUCs ranged from 0.83 to 0.87 (Figure 3). Results from the sensitivity analyses in the Bayesian approach were similar (results not shown).
We also investigated the performance of the different updating methods in the separate operation subgroups. Again, model updating based on 1 year of data led to unstable prognostic factor estimates. This finding was even more pronounced than in the analyses using all cardiac surgery and was especially evident in the AVR and AVR with CABG subgroups (shown for AVR in Figure 4). The use of more data (multiple years) and the Bayesian approach led to more stable estimates of the prognostic factor effects.

**Performance of Updated Models**

Updating the EuroSCORE in the different operation subgroups led to an improved calibration-in-the-large compared with the original EuroSCORE using all methods (AVR results shown in Figure 5). However, when model revision and the closed test procedure were applied to 1 year of data in operation subgroups AVR and AVR with CABG, the calibration slopes were smaller than 1, indicating overfitting. Also, the AUC was lower than that of the original EuroSCORE. This was not the case using other methods or when 3 years of data were used. The addition of operation subgroup as a variable in the EuroSCORE model resulted in better calibration and discrimination than the original model. When compared with model revision in the separate operation subgroups, the model with a variable for operation subgroup resulted in a similar calibration-in-the-large and improved calibration slopes and AUC (Figure 6).

**Benchmarking Using Updated Models**

Updating of the EuroSCORE model led to minor changes in outlier identification (Figure 7). In 2011, Hospital A was identified as an outlier when 3 years of data were used (revision or closed test) or when the Bayesian method was applied. However, it was not considered an outlier using the original EuroSCORE and using 1 year of data (revision or closed test). In 2012, the original model and the Bayesian method resulted in Hospital A to be an outlier.

**Discussion**

**Main Findings**

In this study, 6 methods for model updating were investigated using data on 6 years of cardiac surgery from a large national database. All methods showed an improved calibration-in-the-large compared with the original EuroSCORE, which underlines the need for dynamic modeling. On the other hand, no relevant differences in joint predictor effects (calibration slope) or discrimination (AUC) were identified. This finding could reflect that in our study population, the largest drive behind the calibration drift has been the overall decrease of mortality in time, as opposed to great changes in prognostic factor effects in relation to each other. After all, a simple intercept update and full model revision resulted in comparable calibration slopes and AUCs. The fact that calibration slopes and
AUCs did not improve after updating may also show that the current prognostic factors in the model were not able to do so. This in turn implicates that the residual amount of risk may be attributed to random variation (in a heterogeneous population) or to prognostic factors that are not included in the model.

Additionally, benchmarking results yielded some differences using the different approaches and the original model, but overall, the same centers were identified as outliers. This implicates that in our complete database, all methods are applicable and adequate for updating, and the choice of

**Figure 3.** Validation results of EuroSCORE, updating in all cardiac surgery. All updating methods showed an improved calibration-in-the-large compared with the EuroSCORE. The calibration slope and AUC of the updated models and the EuroSCORE were comparable. AUC indicates area under the receiver operating characteristic curve.

**Figure 4.** Estimated regression coefficients of updated EuroSCORE in operation subgroup AVR. Model updating based on 1 year of data led to unstable prognostic factor estimates, even more pronounced than in the analyses using all cardiac surgery. The use of more data (multiple years) and the Bayesian approach led to more stable estimates of the prognostic factor effects. AVR indicates aortic valve replacement; and LVEF, left ventricular ejection fraction.
which to apply has little clinical implication. It is in smaller databases (eg, in subgroups) that one should be cautious of which approach to use. In smaller subgroups, such as AVR or AVR with CABG, extensive updating, such as re-estimation of all prognostic factors (model revision), led to poor performance when 1 year of data were used. In these subgroups, less extensive updating (recalibration of the intercept and coefficients), more data (a rolling time frame of multiple years), and a Bayesian approach led to better predictions for the subsequent year.

Choice of Method
Our findings are in line with previous research that showed that the preferred method and frequency for model updating depend on the size of the database, the event rate, and the complexity of the model (ie, number of prognostic factors and interaction).15 The larger the database and the event rate, the more extensive updating methods can be adequately performed. The more complex the model, the more data are needed. In model development, a rule of thumb is that a minimum of 10 events is needed for each prognostic factor included in the model. If this rule is applied to model revision, in a situation of 3% mortality and 15 prognostic factors and interaction),15 The larger the database and the event rate, the more extensive updating methods can be adequately performed. The more complex the model, the more data are needed. In model development, a rule of thumb is that a minimum of 10 events is needed for each prognostic factor included in the model. If this rule is applied to model revision, in a situation of 3% mortality and 15 prognostic factors, ≈5000 interventions are needed for complete model revision. In our data, this means that model revision on 1 year of data can only be reliably performed with all cardiac surgery. Although the number of CABG operations was relatively high (over 8000 operations per year), the lower event rate (1.0%–1.5%) limited possibilities for thorough model revision with 1 year of data.

If extensive model updating is applied in smaller databases or in small subgroups, this will result in poor model performance and large fluctuations in the prognostic factor effects over the years. Sudden variation is unlikely and, hence, illustrates that observed noise variation may inappropriate be incorporated in the model. This phenomenon is also referred to as overfitting. When data from 3 years were combined, smoother changes in prognostic factor effects were observed, as well as better predictive performance of the models. Therefore, in case of smaller databases or in small operation subgroups, less extensive updating methods (recalibration) or a rolling time frame of multiple years are recommended.

We illustrated that the inclusion of operation subgroup as a variable in the model may also avert the problem of overfitting, while providing the freedom of modeling varying baseline mortality risks in the different operation subgroups. In our database, the resulting model was superior to the models separately updated within the different subgroups. The effects of all prognostic factors are kept fixed across the operation subgroups, which apparently was a better approach than allowing all effects to vary by subgroup. We did not consider an approach with inclusion of statistical interactions to allow for such variation by subgroup.

The Bayesian method is less familiar than the other frequentist methods but has a strong foundation in statistics.16 In this study, the Bayesian approach resulted in models with good predictive performance. Both in all cardiac surgery, as well as in the smaller subgroups, superior results were found with regard to calibration and discrimination. Therefore, Bayesian model seems to be an excellent alternative for model updating for small as well as large databases.

It is not only the size of the database and the complexity of the risk model that may influence the choice of the updating approach. Thorough analyses of the predictive performance of the risk model may show the most important
weaknesses in its current state. For example, the risk model may strongly overpredict mortality in high-risk patients. If this is the case, the choice for simple intercept update or logistic recalibration is inappropriate and will not lead to better predictions on a patient level because the weighting of the prognostic factors relative to each other remains unchanged.

**Why Dynamic Modeling?**

The patient population in cardiac surgery has changed over the years: patients are becoming older, and an overall worsening of the patient risk profile has been observed in our data, as well as in many other studies. At the same time, the outcomes after cardiac surgery have improved drastically in the last decades. It can hence be expected that the association between patient and intervention characteristics and the studied outcome has also changed in time, the rationale behind new models such as the EuroSCORE II. This theory was recently confirmed by Hickey et al, who showed that the prognostic factor effects of the EuroSCORE model fluctuated over the years in data from the Society of Cardio-Thoracic Surgeons adult cardiac surgery database in Great Britain.

---

**Figure 6.** Calibration-in-the-large, calibration slope, and area under the ROC curve (AUC) of the EuroSCORE, EuroSCORE updated in one operation subgroup, and EuroSCORE updated using all patients with operation subgroup as a variable. The addition of operation subgroup as a variable in the EuroSCORE model resulted in better calibration and discrimination than the original model. When compared with model revision in the separate operation subgroups, the model with a variable for operation subgroup resulted in a comparable calibration-in-the-large (A) and improved calibration slopes (B) and AUC (C). AUC indicates area under the receiver operating characteristic curve; AVR, aortic valve replacement; CABG, coronary artery bypass grafting; and ROC, receiver operating characteristic curve.
Their findings underline the need for dynamic models. Despite these theoretical advantages, this study also shows that benchmarking results in The Netherlands remain unchanged when dynamic models are compared with risk adjustment according to the original EuroSCORE model. This finding does not question the need for dynamic models. First, for patient selection purposes, an accurate estimation of outcomes is desired. Second, over- or underprediction of mortality in specific subgroups may develop in time without immediate effect on the benchmark results if patient populations do not greatly vary across centers, as was the case in our study. However, if it does, a center with a relatively large proportion of a specific group of patient will be benchmarked more favorably (in case of overestimation of mortality) than other centers and benchmarking will be flawed.

Feasibility of Dynamic Modeling

The largest challenge of dynamic modeling is its feasibility. The application of dynamic models is more laborious than that of static models. Some statewide and national registries already use forms of frequent updating, for example, the New York State Cardiac Surgery Reporting System, the Society for Cardiothoracic Surgery in the United Kingdom and Ireland, the Society of Thoracic Surgeons, and the Netherlands Association for Cardio-Thoracic Surgery. The New York State Cardiac Surgery Reporting System fully revises its models every year (model revision); the Society for Cardiothoracic Surgery in the United Kingdom and Ireland and the Netherlands Association for Cardio-Thoracic Surgery calibrate the EuroSCORE to its data with every periodical analysis (recalibration); and the Society of Thoracic Surgeons uses yearly recalibration factors to ensure that the overall observed:expected ratio remains 1. 23–26
Translated into a dynamic modeling approach, the format that is applied by the Society of Thoracic Surgeons is yearly recalibration of a model, combined with a thorough model revision every 3, 4, or 5 years. The updates are primarily aimed at benchmarking of cardiac surgery centers, and it is still the static forms that are mostly used for day-to-day use and patient selection.\(^1,3,24\)

To fully enjoy the advantages of dynamic models, a frequent update to its users is desired. A considerable time lag must always be anticipated with prediction models because data of operations performed today will not be readily available for analyses, and time must pass before outcomes can be assessed (eg, 30-day mortality). Hence, dynamic models require fast data collection, data control, processing, analyses, and publication. A format must be used that ensures periodical uploading of data. Project-based data collection such as used for the development of EuroSCORE and EuroSCORE II is not suitable for dynamic models. In addition, to ensure fast dissemination, online-only publication of models seems to be most appropriate.

**New Prognostic Factors**

The calibration drift that is observed in the EuroSCORE can be attributed to a change in the association between the prognostic factors and the outcome or a change in the incidence of the outcome. Additionally, other prognostic factors may have become increasingly important. This is an issue that has not been discussed in this article. Many other prognostic factors have been mentioned in the literature that may be associated with the risk of mortality after cardiac surgery.\(^27\) Only few of these prognostic factors, such as frailty and preoperative atrial fibrillation, have been shown to indeed improve the predictive performance of existing risk models in the specific population studied.\(^28–30\) However, if predictions remain dissatisfactory after extensive

---

**Figure 6. Continued**

![Graphs showing AUC for CABG, AVR, AVR + CABG, and Other procedures over years from 2008 to 2012. The graphs illustrate the performance of the EuroSCORE, EuroSCORE II, and ReFi in group and all patients.]
model updating, this could indicate that a relevant part of the risk of mortality cannot be attributed to the variables included in the model. In this case, the possibility of new prognostic factors emerging over time should indeed be considered.

**Strengths and Weaknesses of This Study**

This study provides an empirical example of how a static model can be transformed into a dynamic one that can be implemented for benchmarking purposes as well as patient-level information and decision making. The large national database used to perform the analyses was characterized by full nation-wide coverage and a low number of missing values (0.01%). No imputations were needed to perform the analyses. Finally, no new prognostic factors were investigated, which may be interesting for future models.

**Conclusions**

Considering the changing adult cardiac surgery patient population and continuous advances in health care, dynamic models are preferred over static models. This study shows that several methods for dynamic modeling are feasible in a large cardiac surgery database. Resulting models show superior calibration in comparison to the static EuroSCORE model. The preferred method and frequency of model updating depend on the size of the database, the event rate, and the complexity of the model: the more extensive the updating method and the
more complex the risk model, the more events and data are required. Fast data collection, processing, analyses, and publication are essential for dynamic modeling.

Acknowledgments

We thank all the cardiac surgeons in Netherlands for contributing to The Netherlands Association for Cardio-Thoracic Surgery database and all members of the Data Registration Committee for actively participating in the evaluation of data. We also acknowledge the efforts of all people who are involved in data collection and data storage in all the 16 centers: Academic Medical Center (Amsterdam), University Medical Center Utrecht (Utrecht), University Medical Center Groningen (Groningen), Leiden University Medical Center (Leiden), Erasmus University Medical Center (Rotterdam), St Radboud University Medical Center (Nijmegen), VU Medical Center (Amsterdam), Onze Lieve Vrouwe Gasthuis (Amsterdam), Haga Hospital (The Hague), Catharina Hospital (Eindhoven), St Antonius Hospital (Nieuwegein), Maastricht University Medical Center (Maastricht), Isala Klinieken (Zwolle), Medisch Spectrum Twente (Enschede), Medical Center Leeuwarden (Leeuwarden), and Amphia Hospital (Breda). Finally, we thank the department of Medical Informatics at the Academic Medical Center Amsterdam for the storage and processing of data.

Disclosures

None.

References


Improved Prediction by Dynamic Modeling: An Exploratory Study in the Adult Cardiac Surgery Database of the Netherlands Association for Cardio-Thoracic Surgery

Circ Cardiovasc Qual Outcomes. 2016;9:171-181; originally published online March 1, 2016; doi: 10.1161/CIRCOUTCOMES.114.001645
Circulation: Cardiovascular Quality and Outcomes is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2016 American Heart Association, Inc. All rights reserved.
Print ISSN: 1941-7705. Online ISSN: 1941-7713

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circoutcomes.ahajournals.org/content/9/2/171

Data Supplement (unedited) at:
http://circoutcomes.ahajournals.org/content/suppl/2016/03/01/CIRCOUTCOMES.114.001645.DC1

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation: Cardiovascular Quality and Outcomes can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation: Cardiovascular Quality and Outcomes is online at:
http://circoutcomes.ahajournals.org/subscriptions/
Appendix 1

Intercept recalibration, logistic recalibration and model revision

Updating methods ranged from simple recalibration, i.e. re-estimating the model intercept or re-estimating model intercept and calibration slope.\(^{(1,2)}\) to re-estimating individual regression coefficients and the use of shrinkage factors.\(^{(3)}\) The logistic Euroscore is can be calculated by

\[
\text{Predicted mortality} = \frac{e^{\alpha + \sum_{i} \beta_i x_i}}{1 + e^{\alpha + \sum_{i} \beta_i x_i}}
\]

where are \(x_i\) the predictors of the Euroscore, \(\beta_i\) the associated regression coefficients and \(\alpha\) the model intercept. The linear predictor of the Euroscore is given by

\[
Z_0 = \alpha + \sum_{i=1}^{18} \beta_i x_i.
\]

The first two model updating approaches we consider are recalibration methods. The simplest recalibration of a prediction model is re-estimating the model intercept. Re-estimating the model intercept makes sure that the estimates of the prediction model are on average correct. The linear predictor of the model with the updated intercept \((Z_1)\) is equal to

\[
Z_1 = \alpha + Z_0
\]

the parameter \(\alpha\) is estimated by fitting a logistic regression model with only an intercept and an offset variable \(Z_0\). An offset variable is a predictor in a logistic regression model where the coefficient of the predictor is fixed at unity. The second approach, logistic recalibration, updates the model intercept and regression coefficients by a factor that is the same for all coefficients \((\beta_{\text{overall}})\). This is done by fitting a logistic regression model with the linear predictor of the original model as a covariate. The linear predictor of the resulting model is given by

\[
Z_2 = \alpha + \beta_{\text{overall}} Z_0.
\]

Re-estimating the model intercept and logistic recalibration were both applied using one year of patient data to update the Euroscore model.

Another approach to model updating is model revision, which re-estimates individual regression coefficients. The linear predictor of this model is given by

\[
Z_3 = \alpha + \sum_{i=1}^{18} \beta_i x_i.
\]

Model revision requires the re-estimation of a large number of parameters. In small samples the estimates of the regression coefficients can therefore become unstable. This can be remedied by combining multiple years when updating the Euroscore to obtain larger sample sizes or applying shrinkage techniques. We considered using one or three years of patient data to update the Euroscore.

We also applied model revision with shrinkage towards the original regression coefficients of the Euroscore.\(^{(3)}\) Here, again all individual risk factor effects are
updated. After re-estimating the effects shrinkage towards the recalibrated coefficients is applied to stabilize the risk factor estimates. (3) Using these shrinkage techniques the risk factor effects of the updated model will be somewhere between these of model revision and logistic recalibration.

It is also possible to test statistically which update method is most suitable rather than choosing one. The testing procedure we implemented for selecting the most appropriate updating method is a closed testing procedure. The advantage of this procedure is that multiple tests are preformed while maintaining a prespecified statistical significance level alpha (often 0.05). The testing procedure consisted of the following steps:

1. Test the refitted model against the original model, if the refitted model provides a significantly better fit continue, otherwise keep the original model.
2. Test the refitted model against the model with an updated intercept, if the refitted model provides a significantly better fit continue, otherwise use the model with an updated intercept.
3. Test the refitted model against the recalibrated model, if the refitted model provides a significantly better fit use the refitted model, otherwise use the recalibrated model.

All tests were based on the difference in the -2 log-likelihood between the two models. The degrees of freedom in step one of the test procedure was equal to the number of predictors + 1, in step two equal to the number of predictors and in the third step the degrees of freedom was equal to the number of predictors – 1. If the refitted model was chosen by the closed test procedure we applied shrinkage towards the recalibrated regression coefficients of the Euroscore model. The closed test was applied using one or three years of patient data to update the Euroscore model.

**Bayesian learning for updating the Euroscore model**

Updating of prediction models can be considered as a Bayesian process. (4,5) The background knowledge available from the Euroscore, the prior distribution, is combined with the new patient information, the likelihood, to obtain updated estimates of the regression coefficients, the posterior distribution. We used independent normal distributions as the prior distribution of the regression coefficients, to reflect the background knowledge available from the Euroscore model the mean of the normal distributions was equal to the estimated regression coefficients in the Euroscore model. The standard deviation of the normal distributions was set equal to log(4)/2 such that with a probability of 95% the odds ratio of each predictor was between ¼ and 4 times the odds ratio estimated by the Euroscore model. (6) The resulting posterior distribution was approximated using a multivariate normal distribution, this approximation was used as a prior distribution in the next update of the Euroscore.

In Bayesian statistics it is often not possible to analytically derive the posterior distribution, as was the case here. We used Markov Chain Monte Carlo techniques to approximate the posterior distribution. To obtain a prior distribution for the subsequent year we approximated the posterior distribution with a multivariate normal distribution with mean and covariance matrix based
on the samples drawn from the markov chain used to approximate the posterior distribution.
The key issue in the Bayesian approach is selecting the prior distribution. As a sensitivity analysis we reran all analyses using prior distribution with different variances and by first orthogonalizing the predictors in the update set and back-transforming the estimates after updating the EuroSCORE model.