Dynamic Prediction Modeling Approaches for Cardiac Surgery

Graeme L. Hickey, PhD; Stuart W. Grant, MBChB, MRCS; Camila Caiado, PhD; Simon Kendall, MBBS, MS, FRCS (CTh); Joel Dunning, PhD, FRCS (CTh); Michael Poullis, MD, FRCS (CTh); Iain Buchan, MD, FFPH; Ben Bridgewater, PhD, FRCS (CTh)

Background—The calibration of several cardiac clinical prediction models has deteriorated over time. We compare different model fitting approaches for in-hospital mortality after cardiac surgery that adjust for cross-sectional case mix in a heterogeneous patient population.

Methods and Results—Data from >300,000 consecutive cardiac surgery procedures performed at all National Health Service and some private hospitals in England and Wales between April 2001 and March 2011 were extracted from the National Institute for Cardiovascular Outcomes Research clinical registry. The study outcome was in-hospital mortality. Model approaches included not updating, periodic refitting, rolling window, and dynamic logistic regression. Covariate adjustment was made in each model using variables included in the logistic European System for Cardiac Operative Risk Evaluation model. The association between in-hospital mortality and some variables changed with time. Notably, the intercept coefficient has been steadily decreasing during the study period, consistent with decreasing observed mortality. Some risk factors, such as operative urgency and postinfarct ventricular septal defect, have been relatively stable over time, whereas other risk factors, such as left ventricular function and surgery on the thoracic aorta, have been associated with lower risk relative to the static model.

Conclusions—Dynamic models or periodic model refitting is necessary to counteract calibration drift. A dynamic modeling framework that uses contemporary and available historic data can provide a continuously smooth update mechanism that also allows for inferences to be made on individual risk factors. Better models that withstand the effects of time give advantages for governance, quality improvement, and patient-level decision making. (Circ Cardiovasc Qual Outcomes. 2013;6:649-658.)

Key Words: Bayesian forecast • calibration • clinical governance • logistic model

Clinical prediction models (CPMs) have been routinely used in the field of cardiac surgery for the past 2 decades. They have been used in the context of decision making, governance analyses, and patient assignment in randomized controlled trials. CPMs for decision making allow both patients and surgeons to weigh the risks of undergoing surgery. For the purposes of governance, CPMs are used to adjust observed outcomes for the heterogeneous case mix in the patient population, thus allowing for comparison between healthcare providers. The assignment of patients to different treatment arms in randomized control trials has also been based on scoring through CPMs. In all 3 applications, it is clear that there is an overwhelming requirement to use accurate CPMs. The repercussions of using an inadequate or outdated CPM can be important: patients and clinicians can make suboptimal decisions, for example, denying an apparent high-risk patient potentially life-saving surgery; healthcare providers can be falsely identified as poorly performing outliers or falsely assured that their performance is satisfactory; and clinical trial design protocols can be inadvertently affected.

In the United Kingdom, the European System for Cardiac Operative Risk Evaluation (EuroSCORE) model has been the predominant CPM used in all contexts. It was originally developed on a clinical dataset prospectively collected in 1995. Despite its ability to discriminate between low- and high-risk patients remaining consistently good during the course of its use, it has now been widely found to overpredict the risk of mortality for contemporary cardiac surgery. This loss of calibration, the agreement between observed and predicted event rates, has been found to have occurred over...
WHAT IS KNOWN

• Clinical risk prediction models, such as the Society of Thoracic Surgeons or European System for Cardiac Operative Risk Evaluation, have found widespread use in cardiac surgery for quality improvement, governance, patient-level decision support, and other academic purposes.
• As quality of care has improved over time, these models have been prone to calibration drift, which needs to be acknowledged and addressed, or else risk models may become dangerous for patients.

WHAT THE STUDY ADDS

• Here, Hickey et al explore possible model refit methods for updating risk prediction models, based on the European System for Cardiac Operative Risk Evaluation model and its covariates, and compare the results against the static approach.
• Increasingly sophisticated approaches lead to better prediction, with a potential ability to predict for the future. However, the increasing complexity of the methods needs to be balanced against their potential for clinical use and feasibility to implement.
• The findings have implications for cardiac surgery risk prediction, but there are wider implications as clinical registries, decision support tools, and quality improvement programs become more widely implemented across medicine.

Methods

Data Extraction and Cleaning
Before the analysis, in Great Britain and Ireland all prospectively collected Society for Cardiothoracic Surgery data were extracted from the National Institute for Cardiovascular Outcomes Research (NICOR) clinical registry for all adult cardiac surgery procedures performed in National Health Service hospitals and some private hospitals in England and Wales between April 1, 2001, and March 31, 2011. A reproducible cleaning program was developed and applied to the data before any analysis to resolve transcriptional discrepancies, numeric irregularities, and clinical data conflicts. Duplicate records and records with incoherent date stamps were also removed by the data cleaning macros.20 Variable definitions for the study are available at http://www.ucl.ac.uk/nicor/audits/Adultcardiacsurgery/datasets. Using a purpose-built algorithm that takes into account data conflicts and missingness, records that did not correspond to either the first or only cardiac procedure within a single admission spell were removed to prevent mortality being double-counted for a procedure. After data cleaning, summaries of the most recent 3-year database subset were returned to each contributing hospital for local validation. Previous data were validated as part of the ongoing National Adult Cardiac Surgery Audit program.

The outcome for the study was in-hospital mortality, defined as death because of any cause during admission at the hospital where the cardiac surgery was performed. For records in which the hospital discharge status was unknown or in doubt, record linkage to an Office for National Statistics (UK) database, which records details of all deaths in the United Kingdom, was used to backfill the database. Records were excluded from the analysis if outcome data were not available after cleaning. All transplant, trauma, or primary ventricular assist device procedures were also excluded following Society for Cardiothoracic Surgery in Great Britain and Ireland governance protocols. This study was approved by the NICOR National Adult Cardiac Surgery Audit Research Group, part of University College London. Patient consent was not required because data were pseudonymized at the point of extraction.

Missing Data and Imputation
In the interests of comparison, we consider only model adjustment for the 17 variables in the logistic EuroSCORE model (see Table for a list of variables and their clinical definitions). If a categorical/binary EuroSCORE variable was missing for a record, then it was imputed by the reference value. For example, if no data were recorded about surgery on the thoracic aorta, then it was assumed that the patient did not have surgery on the thoracic aorta. This assumption has been robust to validation exercises and is based on expert understanding of the clinical data-collection process. The percentage of missing data for each variable has been previously described.13 For 9 records in which the patient age at the time of surgery was missing, it was imputed with the median age for the associated financial year.

Data-Collection Mechanism
Data are continuously being collected by UK hospitals and uploaded to the NICOR clinical registry as part of the ongoing National Adult Cardiac Surgery Audit program. Data are validated by individual hospitals, NICOR, and analysts on a regular basis. Upload to the database varies according to hospital; some hospitals enter data regularly, whereas other hospitals upload large batches periodically. For the purposes of this analysis, we will assume that the data are uploaded each calendar month and are available for analysis immediately; this is in line with future plans for national data collection. However, this interval could be adjusted without loss of generality. We also assume that for each monthly batch of data, the individual observations in the subset are independent; that is, the time of day and day of month are irrelevant for modeling.
Four different approaches are described and fitted to model the cross-sectional risk of outcome using a standard multiple logistic regression model with the logistic EuroSCORE variables. The first approach represents the status quo, which is nondynamic, and approaches 2 to 4 are considered as dynamic alternatives.

**Approach 1: No Updating**

A standard CPM is fit using multiple logistic regression to the first 12 months of data (24,180 records; this is comparable to the 19,030 records used to fit the original EuroSCORE model). For the following 9 years of the study period, the model is not updated; hence, the model coefficients are constant during the entire 10-year period. Note that we have not used the original EuroSCORE model coefficients so that all models are contemporary, and thus comparable, at the start of the study period.

**Approach 2: Periodic Update**

Anticipating a loss of predictive model calibration, namely, the calibration outside of the data used to fit (or train) the model, the CPM is updated (refit) periodically. For purposes of illustration, we consider 2 periods: every 12 months and every 24 months. The first example yields 10 sets of model coefficients during the study period, and the second example yields 5 sets of model coefficients. Note that during the first 12-month study period, the coefficient estimates of approach 2 (12 monthly update) will coincide with the estimates determined from approach 1.

**Approach 3: Rolling Window**

To stabilize coefficient estimates and avoid large abrupt changes in the associated risk with each model variable between the change points of approach 2, we periodically refit the CPM using a rolling 24-month window. For the purposes of illustration, we consider 2 options for the frequency to re-estimate model coefficients: every 12 months and every 1 month. For example, in the first illustration, we first fit the model using data from April 1, 2001, to March 31, 2003. Then we estimate the coefficients for the period April 1, 2003, to March 31, 2004 (12 months), using the model fitted to the data for April 1, 2002, to March 31, 2004 (24 months). This method is iterated until the end of the study period. The reason for considering monthly updates is for comparison with approach 4. Note that in this example, during the first 24 months of the study period, the coefficient estimates of approach 3 will coincide with the estimates determined in approach 2 (24 monthly update) for the same period. They will also coincide every alternating year.

**Approach 4: Dynamic Logistic Regression**

Starting from a CPM fitted to 12 months of training data using the same methods as per approach 1, a dynamic logistic regression model is formulated and fitted within the Bayesian modeling paradigm to update the model coefficient estimates each month thereafter. This
Bayesian forecasting model and proposed estimation procedure are described by McCormick et al., and technical details are provided in the Appendix in the online-only Data Supplement. Briefly, the model coefficients estimated at time $t$ are used to construct a prior distribution for the coefficients at time $t+1$. This is done by first constructing a prediction equation using the data up until time $t$ and then combining it with the observed data at time $t+1$ to yield an updating equation, which admits the coefficient estimates. Simply, we are using the approximated posterior distribution at time $t$ as the prior distribution for time $t+1$ in the prediction equation. Inherent to this modeling approach is a vector of dynamic tuning parameters (1 per estimated model coefficient), which effectively determines how much weight is given to historical data. Following McCormick et al., we a priori opted to dichotomize the tuning parameters and constrain the model to only estimate 5 at each time point, bringing the number of calculations down from 524 288 to a more manageable 32. The values of the dichotomization were chosen after post hoc analysis.

For each risk factor, we plot as time series the changes in coefficient estimates (solid lines) and an $\pm$95% pointwise confidence interval band (dashed lines). This allows for inferences to be made about the relative degree of association between the outcome and risk factor after adjusting for the other variables. For approach 1, we omit the confidence interval estimate because the uncertainty only pertains to the initial training phase; the estimated coefficients are only fixed efficient estimates (solid lines) and an 95% pointwise confidence interval (dashed lines). This allows for inferences to be made.

### Results

Of 318 681 cleaned records extracted for the time period and regions, 1300 were removed because they represented within admission re-do procedures. A further 584 records were removed based on procedure exclusion criteria, and 165 records were removed because of missing outcome data. In total, 316 632 records were included in the study from 37 different hospitals. Line graphs showing the frequency of records for each month and the number of contributing hospitals for each financial year to the study are shown in Figure 1. The first and second years of data contained 24 180 and 30 241 records, respectively, and were set aside for model training in the various approaches. The model fits during the training phase of approaches 1, 2 (12-month periodic refit) and 4 and are identical. After the post hoc analysis comparing approach 4 to approaches 1 to 3 with different tuning parameters, a suitable compromise between reactivity of the model estimates to changes and smoothness was identified (see online-only Data Supplement).

Figure 2 plots the time series for the model intercept coefficient based on each of the approaches. Similarly, Figures 3 to 5 plot the time series for the model coefficients for 6 of the following variables: sex, chronic pulmonary disease, neurological dysfunction, poor left ventricular function, recent myocardial infarction, and postinfarct ventricular septal defect. The remaining 12 EuroSCORE coefficients are shown in the online-only Data Supplement.

The decreasing intercept coefficients as calculated by approaches 2 to 4 are consistent with the observed continual reduction of in-hospital mortality rates in England and Wales during this period. The analysis of the patient-related factors yields several interesting points. Women have been associated with increasing risk of in-hospital mortality that peaked at the midpoint of the study period. The risk associated with both extracardiac arteriopathy and chronic pulmonary disease became relatively stable during the study period after a period of initial adjustment. With the exception of a few periods, including the training period, the coefficient for neurological dysfunction was not statistically significant during the study period at the 2-tailed 5% significance level, even yielding counterintuitive inferences (ie, negative coefficient estimates) at times. Cardiac-related risk factors have also yielded insights during the study period. For example, the association of patients with poor left ventricular function and model outcome has been decreasing year-on-year and now seems to be plateauing. The coefficient for recent myocardial infarction has increased to almost twice that inferred from the training period. Post infarct ventricular septal defect has remained relatively steady during the entire study period. Similarly, interesting trends can be found in the other 12 EuroSCORE coefficients (see online-only Data Supplement).

The 2 approaches that yielded monthly coefficient updates (approach 3 [1 month] and approach 4) tracked comparably throughout the study period, each giving smooth time series. Only in cases of abrupt changes did the 2 approaches noticeably diverge; however, these separations were small. The approaches with time coarsened to 12 months were quite often subject to abrupt changes, in particular approach 2 (12 month), which had no historical data included to stabilize estimates. This led to scenarios whereby log-odds ratios were halved or doubled going across the financial year threshold. Approaches 2 (24 month) and 3 (12 month), in contrast, yielded coefficient...
estimates with relatively smaller variability overall compared with approach 2 (12 month).

**Discussion**

**Summary of Model Approaches**

In this study, 3 approaches have been analyzed for fitting CPMs to data with a nonstationary data-generating process and juxtaposed to a static model approach. Approach 1, the status quo, was developed on a snapshot of data for a population and continually used thereafter without further update. There has been prolific use of such CPMs in cardiac surgery, the Parsonnet score and EuroSCORE being 2 of the most well known. The EuroSCORE is still actively recommended by UK health-care commissioners for determining whether a patient should receive a transcatheter aortic valve implantation or surgical aortic valve replacement for the treatment of aortic valve stenosis and is one of the inclusion criteria for the Placement of Aortic Transcatheter Valves (PARTNER) trial.\(^5\)\(^,\)\(^24\) It has already been shown that such models have a limited shelf-life and are not appropriate for long-term use without being updated.\(^13\) Models that systematically lose calibration over time are intrinsically dangerous because they provide misleading indications of risk to support patient-level decision making and can provide false reassurance to providers about quality of care.\(^6\)

Approach 2 reflects practice by 2 major governance programs run by professional cardiac surgery societies and also addresses the call for periodic recalibration.\(^25\) However, the discontinuity of the risk factor coefficients between the update time boundaries is arbitrary and reminiscent of the statistical debate on dichotomization of continuous risk factors.\(^26\) Sudden changes in the distribution of risk factors can lead to abrupt changes in the discontinuous model coefficient estimates. In any given discrete period of time, these changes could be an artifact of extraneous variation such as data quality or low prevalence of the risk factor; hence, the model might be unsuitable for application. Approach 3 is a natural extension to approach 2, which dampens the abrupt changes by drawing on previous data, thus stabilizing coefficient estimates.

The final approach discussed, approach 4, describes a more formal approach that acknowledges the hypothesized dynamic process. Although not shown here, approach 4 could be exploited to make predictions about the model parameters in future periods, taking into account the current trajectory and uncertainty. The other models, however, would only extrapolate from the last known coefficient estimates without further adjustment. Both approaches with monthly coefficient updates yielded similar inferences. There is a degree of arbitrariness in the selection of the rolling window and tuning parameters for approaches 3 and 4, respectively. Nonetheless, the tuning parameters in approach 4 can be modeled in the Bayesian framework, and estimators can be determined sequentially conditional on the observed data. This would increase the computational resources required, and therefore
approximations would have to be developed. Moreover, although approach 3 discards data outside of its rolling window, approach 4 effectively uses all data using weights to give greater importance to more recent data. The combination of these features makes approach 4 more appealing, especially when some coefficients might be relatively volatile because of a low number of events.

In-hospital mortality rates have improved, despite an increasingly high-risk patient population, and this is reflected in the downward trend of the model intercept coefficient for the periodically refitted and dynamic models from a baseline measure defined by a historical static model. Potential factors contributing to the inherent dynamics in observed mortality include improving standards in surgery and postoperative care, changes to local and national healthcare policy, and the feedback of monitoring data to drive improvements in quality. Simultaneously, the expected mortality is subject to dynamic influences. These include changes to the population of patients coming to surgery, possible earlier detection of serious complications, and changes to the definition of risk factors.

Strengths and Weaknesses Compared With Other Studies
Steyerberg proposed a multitier structure for updating CPMs as new data comes in, starting with (1) recalibration of the model intercept, (2) model recalibration of the slope and intercept parameter, (3) model refitting, and (4) model extension with new covariates. Notwithstanding the simplicity of steps 1 and 2 in Steyerberg’s hierarchy, they do not account...
for changes between variables. Steps 1 and 2 are likely to be valuable for CPMs where data are limited; however, a data-rich registry, such as the NICOR or the Society of Thoracic Surgeons database, is ideally suited for a model refit approach. Approach 2 is equivalent to step 3 in Steyerberg’s hierarchy. Implementation of step 4 would be time consuming and is not something that could be fully automated, because in general, model building requires careful analysis. A proposal on when to move between each step based on a likelihood ratio test was suggested but still remains ad hoc.

Previous studies have also considered dynamic generalized linear model approaches. However, unlike the study here that modeled the changes in coefficients, the expected outcome was directly modeled. In addition, other studies have assumed that the risk score was known; this was (1) because of potential sensitivity in the estimation method to outliers, and (2) because degrading model performance may be attributable to the choice of risk factors rather than the coefficient value. We agree that the model will probably be sensitive to the choice of which covariates to include, but this assumption might be a limitation for prolonged study periods such as the one considered here. Generally, however, other methodological proposals for detecting outlier healthcare providers that are based on sequential historical series will be limited by the assumption of model stationarity.

Limitations in Using Dynamic Models for Clinical Prediction
There are limitations in using dynamical modeling approaches for clinical prediction: (1) increased complexity in model
(2) accounting for emerging risk factors, (3) increased difficulty in summarizing model performance, (4) difficulties in making models transparent, (5) application of the models in practice and the need for regular data expansion, (6) problems with acceptance of the models by healthcare providers because of the increased complexity of the methodology used for model development, and (7) lags in collecting new data.

Developments in informatics, especially online tools, mean that routine use of dynamic models is a challenge that can potentially be overcome. However, with regard to approach 4, the need to specify tuning parameters is difficult, although with sufficient computational resources, this could be continuously optimized. Emerging risk factors are also a potential limitation of static CPMs (ie, those that are never updated) and will likely influence routine model-selection procedures during the development of CPMs. The standard means of studying model performance, namely, through measuring model calibration and discrimination, are extendable to the dynamic model approaches here. However, it must be recognized that model performance will also be dynamic, and a model averaging approach may be advantageous. It is important that a CPM has clinical validity. Despite the added complexity and in light of recent evidence showing the systematic loss of calibration for EuroSCORE, dynamic models should satisfy the face validity requirement.

Dynamic models pose a wider challenge to those who use them and those who would govern them. The current system of publishing a CPM in a journal article (or otherwise) is not tenable here; updates would need to be made, and users would have to be aware of this. Along with the need
to update models drives a requirement for regular and timely data submission. Such an approach would need centralized management and coordination. Professional societies, such as the Society for Cardiothoracic Surgery in Great Britain and Ireland and Society of Thoracic Surgeons, would be likely to play a key role in the coordination of this task. It is worth noting that dynamic models could be run online with limited or no prognostic data held in memory, which is advantageous when personal or identifiable patient data are to be used. The adaptive nature inherent to dynamic modeling approaches may reduce the appropriateness of using the models on different patient populations; however, this potential limitation would need to be studied further. The outcome for this study was in-hospital mortality, which is typically known by the clinician soon after the event; hence, the data collection will be contemporary. In other cases, for example, where the outcome is 1-year mortality, there will be an unavoidable lag between the data and analysis.

Limitations of This Study
As for all research projects based on clinical registry data, data quality is a limitation of this study. In particular, the methodology used to impute the data may be a source of statistical bias, whereas more sophisticated methods such as multiple imputation would be preferable.33 However, missingness was generally low and more robust methodology, such as multiple imputations, that would serve only to complicate matters, especially if there is a dynamic structure to the data-generating process. Database errors have been demonstrated to affect mortality rate estimates,34; however, record linkage to Office for National Statistics data and validation exercises have confirmed accuracy of the outcome data for this study. A further limitation of this study was that model validation was not undertaken. This is an important aspect of model development; however, a dynamical model framework increases the complexity of this enormously. Nonetheless, dynamically monitoring the model discrimination would provide a useful measure, but this is beyond the scope of this study.

All approaches have a degree of arbitrariness about them with regard to the (1) initial training period, and (2) update frequency, and these might lead to different results. We demonstrated multiple examples of approaches 2 and 3 to highlight this. Finally, we did not develop a model; instead, we opted to focus on the EuroSCORE variables because it is the motivating example for this research. In principle, if one were to set up a dynamic CPM for use, one would follow common practice; the Society of Thoracic Surgeons risk models are a good example of this.36 Because of the issues discussed, such as emerging risk factors, to sensibly automate this in a dynamic framework would require the use of a model-averaging approach. This is beyond the scope of this research; however, McCormick et al described how to embed this into approach 4.31

Meaning of Study
The use of the original EuroSCORE is an extreme example in the cardiac surgery specialism but an important one nevertheless. Although it has been recently shown that EuroSCORE II is a well-calibrated model for contemporary cardiac surgery in the United Kingdom,37,38 there is no reason to suggest that its calibration will not deteriorate in the future as has been observed for the original EuroSCORE models. Moreover, a future one-off recalibration or model update may only remediate the issue in the short term. The same argument also extends to other routinely used cardiac surgery CPMs, such as the Society of Thoracic Surgeons models.39,40 If the dynamics inherent to cardiac surgery, especially the baseline risk, continue to change, then calibration drift is inevitable. Dynamic models or periodic model refitting is a necessary approach to counteract this calibration drift.

Acknowledgments
The National Institute for Cardiovascular Outcomes Research, University College London, London, provided the data for this study.

Sources of Funding
This research was funded by a grant from Heart Research UK (grant no. RG2583).

Disclosures
None.

References


Dynamic Prediction Modeling Approaches for Cardiac Surgery
Graeme L. Hickey, Stuart W. Grant, Camila Caiado, Simon Kendall, Joel Dunning, Michael Poullis, Iain Buchan and Ben Bridgewater

_Circ Cardiovasc Qual Outcomes_. 2013;6:649-658; originally published online October 22, 2013;
doi: 10.1161/CIRCOUTCOMES.111.000012
_Circulation: Cardiovascular Quality and Outcomes_ is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2013 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/6/6/649

Data Supplement (unedited) at:
http://circoutcomes.ahajournals.org/content/suppl/2013/10/22/CIRCOUTCOMES.111.000012.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/
SUPPLEMENTAL MATERIAL

Appendix: Dynamic Prediction Modelling Approaches for Cardiac Surgery

Data

We envisage a hypothetical system in the form of a live sequentially expanding database made up of $Y^t = [Y^{(1)}, Y^{(2)}, \ldots, Y^{(t)}]^T$ and $X^t = [X^{(1)}, X^{(2)}, \ldots, X^{(t)}]^T$, each $Y^{(t)}$ and $X^{(t)}$ represents the vector of binary outcomes (deceased/alive) and corresponding design matrix of measured EuroSCORE risk factor data (columns) for all records (rows) during time period $t$ respectively. The superscript $T$ denotes the mathematical matrix transpose notation. We assume in any given time period $t$ the observations in each data subset $(Y^{(t)}, X^{(t)})$ are considered independent; that is the time of day and day of month are irrelevant for modelling. For this analysis we assume each time interval to be one calendar month from here onwards.

Logistic regression model

For all approaches described in this research, we assume the logistic regression model cross-sectional for the appropriate time window. For some given time window the model specifies that for data record $i$, the binary outcome $Y_i$ has a Bernoulli$(p_i)$ distribution, where $p_i$ is the probability of experiencing an in-hospital death, such that

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{18} X_{18i},$$

and $X_{1i}, X_{2i}, \ldots, X_{18i}$ denote the 18 measured EuroSCORE variables for patient $i$, and $\beta_0, \beta_1, \ldots, \beta_{18}$ are the model coefficients to estimate (with $\beta_0$ as the intercept coefficient). Note that the left ventricular function variable, which is categorical, requires two coefficients following transformation using dummy variables.

Approach 1

Approach 1 is a proxy for the ubiquitous logistic EuroSCORE model. Standard logistic regression model fitting software is used to fit the model to the subset of first year of data $(Y^{1:12}, X^{1:12})$ to generate coefficient estimates and approximate 95% confidence intervals. The coefficient estimates are then extrapolated for the following 9-years, i.e. time $t = 13$ months to $t = 120$ months.
Approach 2

Approach 2 is a quasi-dynamic model that incorporates contemporary adjustments to a limited degree by independently iterating Approach 1 periodically. It assumes piecewise model stationarity and can be considered as a special case of a change-point model. In the case of 12-monthly updates, it is fit similarly to Approach 1, except further models are fit independently to the data \((Y_{13:24}, X_{13:24}); (Y_{25:36}, X_{25:36}); \) etc. Thus for each 12-month interval, separate estimated model coefficients are determined. The principle is the same for the 2-year windows: models are independently fit to \((Y_{1:24}, X_{1:24}); (Y_{25:48}, X_{25:48}); \) etc.

Approach 3

Approach 3 uses a progressively rolling fixed-width window, which can be considered an extension of Approach 2. Starting from an initial window of 24-months, a model is fit according to the standard Approach 1, thus yielding coefficient estimates for \(t = 1\) to \(t = 24\). After this the 24-month window is incrementally moved forward, say, 12-months. The data \((Y_{13:36}, X_{13:36})\) is then used to estimate the coefficients for the time \(t = 25\) to \(t = 36\) (inclusive). This procedure is continued estimating coefficients for each 12-month block. The principal is the same for the 1-month increment, yielding 97 sets of coefficient estimates as opposed to 9 for the 12-month increment.

Approach 4

Approach 4 is a Bayesian forecasting method that is relatively more complicated to fit and is described in full detail by McCormick et al. We briefly outline their method here. First, a state-space equation for the model coefficients is proposed:

\[
\beta^{(t)} = \beta^{(t-1)} + \delta^{(t)},
\]

where \(\beta^{(t)}\) denotes the vector of model coefficients at time \(t\), i.e. \(\beta^{(t)} = (\beta_0^{(t)}, \beta_1^{(t)}, ..., \beta_W^{(t)})\), and \(\delta^{(t)}\) is a random vector drawn from a multivariate normal distribution with mean 0 and covariance matrix \(W_r\).

Following our notation, the recursive estimation procedure at time \(t > 12\) is then:

1. Assume that \(\beta^{(t-1)} | (Y^{(t-1)}, X^{(t-1)}) \sim N(\tilde{\beta}^{(t-1)}, \Sigma^{(t-1)})\) for some suitable values of mean \(\tilde{\beta}^{(t-1)}\) and covariance matrix \(\Sigma^{(t-1)}\), where \(N(\cdot, \cdot)\) indicates a (multivariate) normal distribution. The model is initialized by fitting the clinical prediction model to the training
data \((Y_{1:12}, X_{1:12})\) as per Approach 1 and using the corresponding estimates to specify the mean and covariance matrix. The coefficient estimates of Approaches 1 and 4 coincide for the first 12-months therefore.

2. From the state-space equation and the assumption in step 1, we obtain the prediction equation

\[
\beta^{(t)} \mid (Y^{(t-1)}, X^{(t-1)}) \sim N(\beta^{(t-1)}, \Sigma^{(t-1)} + W_t),
\]

To simplify the model a strong assumption that \(W_t\) is proportional to \(\Sigma^{(t-1)}\) is made such that

\[
\beta^{(t)} \mid (Y^{(t-1)}, X^{(t-1)}) \sim N(\beta^{(t-1)}, \Sigma^{(t-1)} / \lambda_t),
\]

where \(\lambda_t\) is a tuning parameter, which can either be a constant or a vector of tuning parameters having the same dimension as \(\beta^0\).

3. After observing the data at time \(t\), the posterior distribution of \(\beta^{(t)}\) is proportional to the likelihood function for \(\beta^0\) conditional on \((Y^0, X^0)\) (which immediately follows from the logistic regression model) multiplied by the prediction equation. As this posterior distribution is not available in closed form, a normal approximation is used to yield joint modal estimates, \(\hat{\beta}^{(t)}\) and \(\hat{\Sigma}^{(t)}\) - the updating equation(s). See Equations 4-6 in McCormick, Raftery, Madigan, et al. (2012) for exact equations.

4. Steps 1 – 3 are repeated thereafter for the next time period.

Forgetting parameter

The tuning parameter, \(\lambda_t\), is selected to attain a fixed degree of ‘forgetting’; McCormick et al. refer to this as a ‘forgetting factor’ and proposed that, in the interests of computational efficiency, two values of \(\lambda_t\) are compared at any one point: “no forgetting” \((\lambda_t = 1)\) and “some forgetting” \((\lambda_t = c)\) for some constant \(c < 1\). At each step, the value that maximises the predictive likelihood is selected. The subscript \(t\) indicates that the forgetting factor can be tuned at each separate time period; however this is computationally expensive to do for all \(2^{19}\) possible configurations (i.e. 2 tuning parameter values and 19 model coefficients).

Therefore we adopted the strategy of optimizing between all combinations of “not-forgetting” \((\lambda_t = 1)\) versus “some-forgetting” \((\lambda_t = c)\) for four logistic EuroSCORE variables considered to be potentially influential \textit{a priori} in the calibration drift: the intercept, patient age at time of surgery, ‘surgery other than isolated coronary artery bypass grafting’ and ‘surgery on the thoracic aorta’. For all other risk factors in the model, we constrained them to share the same forgetting factor at each time \(t\), which were simultaneously optimized between \(\lambda_t = 1\) and \(\lambda_t = c\). It has been noted that the forgetting factor can be interpreted in terms of the data being weighted in the estimation procedure with exponentially
decaying weights. McCormick et al. proposed setting $c = 0.99$ following an investigation into the sensitivity of the parameter on inferences. Our analysis (results not shown) found that this choice reduced the ability of the coefficients to react to increased acceleration of changes in association with the outcome. Through evaluation of a number of values $c = 0.999$, 0.99, 0.90, 0.80 we concluded that $c = 0.90$ was sufficient to capture this data whilst yielding a smooth model fit. Recommendations on how one might optimise the choice of $c$ in the future are given in McCormick et al. These include using a grid of values; using Bayes factors to evaluate whether the tuning parameters are dynamic; prior selection of values based on prior expert judgements of system stability; and factoring in the availability of computational resources to necessitate updates.

**Software**

Approaches 1 to 3 were fitted in R (version 2.15.2) using the standard glm function for fitting generalised linear regression models. Approach 4 was fitted in R using scripts specifically written for this analysis that exploit the functions in the R dma package.

**Comparison of approaches**

Figures 2-5 in the main manuscript describe the different approaches for the intercept parameter and six of the EuroSCORE variables. The remaining 12 EuroSCORE variables are shown in Figures A1-A2 below.

**References**


**Figure A1.** Model coefficients for EuroSCORE variables over time for each model. Solid lines correspond to estimates; dashed lines correspond to 95% confidence intervals. Abbreviations: Pre-op – preoperative state.
Figure A2. Model coefficients for EuroSCORE variables over time for each model. Solid lines correspond to estimates; dashed lines correspond to 95% confidence intervals. Abbreviations: LV – left ventricular; CABG – coronary artery bypass graft.