View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of Liverpool Repository

A review of statistical updating methods for clinical prediction models

Authors: Su, Ting-Li; Jaki, Thomas; Hickey, Graeme L; Buchan, Iain; Sperrin, Matthew.

Abstract

A clinical prediction model (CPM) is a tool for predicting healthcare outcomes, usually within a specific population and context. A common approach is to develop a new CPM for each population and context, however, this wastes potentially useful historical information. A better approach is to update or incorporate the existing CPMs already developed for use in similar contexts or populations. In addition, CPMs commonly become miscalibrated over time, and need replacing or updating. In this paper we review a range of approaches for re-using and updating CPMs; these fall in three main categories: simple coefficient updating; combining multiple previous CPMs in a meta-model; and dynamic updating of models. We evaluated the performance (discrimination and calibration) of the different strategies using data on mortality following cardiac surgery in the UK: We found that no single strategy performed sufficiently well to be used to the exclusion of the others. In conclusion, useful tools exist for updating existing CPMs to a new population or context, and these should be implemented rather than developing a new CPM from scratch, using a breadth of complementary statistical methods.

Keywords: clinical prediction model; model updating; model validation; model recalibration; risk score

1. Introduction

Clinical prediction models (CPMs) are tools for predicting the natural course of diseases or the responses of patients to healthcare interventions, with regard to specific endpoints and observable characteristics[1]. For example, clinicians, healthcare managers and patients may be interested in assessing the risk of dying within 30 days of undergoing a heart bypass operation. We expect this risk to depend both on the characteristics of the patient, such as gender, age, and comorbidities, and on the characteristics of the intervention, such as the experience of the surgeon. A CPM is usually developed by fitting a statistical model to existing data. The choice of model to be fitted depends on the nature of the endpoint; common choices are logistic regression (for a binary endpoint) and survival models (for a time-to-event endpoint).

CPMs have three main practical uses. First, they may be used at an individual patient level to communicate risk and aid in the clinical decision-making process by stratifying patients into different treatment option groups[2] or to determine whether further testing is warranted to reach an appropriate decision[3,4]. Second, they may be used for planning healthcare services by predicting disease prevalence and future demand on services, or to explore the consequences of different local policy options. Third, they may be used in the quality management of healthcare services, where clinical audit processes compare observed with expected outcomes, given appropriate adjustments for differences in case-mix (e.g. ensuring the surgeon who takes

on difficult cases, with a higher baseline risk, is appropriately compared with his/her peers who operate on lower risk patients)[5,6].

The topic of developing, validating and using CPMs receives considerable attention in the statistical and clinical literature; for a recent overview see the PROGRESS paper series [2,7–9]. The importance of transparent reporting of the development, monitoring and validation of CPMs has recently been emphasised by the TRIPOD statement[10].

In practice, CPMs are usually selected or developed for a given population and endpoint of interest. There are two general approaches: 1) develop a new CPM in the population of interest; or 2) use an existing CPM that has been developed and used in related contexts. The first approach wastes prior information, risks over-fitting, and ultimately leads to many CPMs existing for the same endpoint, which is confusing and makes it difficult to decide which one to apply in practice. The second approach may result in a CPM that is not fit for purpose, poorly calibrated and lacking discrimination. A better way forward may be to combine these approaches and work from the 'middle ground' in which existing CPMs that may be relevant for the population and endpoint of interest are taken, and revised to suit the new population.

Another common pitfall with CPMs is that their performance can deteriorate over time: calibration drift[11](P392). This can be attributed to changes over time in: prevailing disease risks (e.g. the obesity epidemic accelerating the force of diabetes morbidity); unmeasured risk factors for disease and treatment outcomes; treatments; treatment settings; adjunct

treatments and wider healthcare; and data quality. Therefore, to remain valid, CPMs must evolve over time – either by renewing or updating the model at discrete timepoints[12], or by allowing the CPM to operate dynamically, updating continuously in an online fashion[13].

The quantitative performance of a CPM can be evaluated through its discrimination (how well patients with poor outcomes are separated from those with better outcomes) and calibration (agreement between probabilities from the CPM and observed outcome proportions). These can be assessed internally (using, for example, cross-validation to correct for within-sample optimism) or, more preferably, externally using a different population[14]. The discrimination is measured by the area under the receiver operating characteristic (ROC) curve (AUC)[15], with a larger AUC indicating a better prediction model. The ROC is a plot of the sensitivity versus 1-specificity for a CPM, based on dichotomizing the predicted probabilities from this CPM into disease and non-disease two groups over a continuous range of thresholds. Approaches exist to construct a 95% confidence interval (CI) for the AUC[16–19]. A calibration plot[20] plots the observed against the predicted outcome probabilities. For a perfectly calibrated model, this should fall on a 45 degree straight line. A univariate logistic regression model can also be used to assess calibration: a calibration intercept can be obtained by regressing the binary outcome on the predicted log-odds while fixing this log-odds as an offset variable ; and a calibration slope can be obtained by a separate fit while fixing the intercept at previously estimated value[21]. A method with good calibration should have zero intercept and a slope of one. If the intercept is greater (smaller) than zero, it indicates the prediction is systematically too small

(too large); if slope is bigger (smaller) than one, it indicates CPM is under-fitting (over-fitting) the data. Other measures to assess the performance of CPMs can be found in Steyerberg et al.[4] or Austin et al[22]. A model with poor calibration can be recalibrated easily, whereas poor discrimination is far more difficult to improve.

The aim of this paper is to highlight and compare various statistical strategies for modifying existing CPMs to perform well in a new population, and strategies to maintain performance over time. In Section 2 we review various statistical strategies for updating existing CPMs, focussing on developments since 2004, but with historical references where appropriate. We illustrate the application of these strategies using data from the UK and Ireland National Adult Cardiac Surgery Audit (NACSA) registry cardiac surgery data, which is introduced in Section 3. The performance of the selected strategies as applied to the NACSA data are presented in Section 4. We conclude in Section 5 with a discussion.

2. Methods for updating CPMs

Focussing on the statistical literature of the last 12 years (2004-2015), we have identified three main approaches for updating CPMs in light of new data. The first approach, which we term *regression coefficients updating*, focuses on updating some or all coefficients from an existing CPM. The second approach is *meta-model updating*, which synchronizes multiple existing CPMs into one new meta-CPM. The third approach is *dynamic updating*, in which one or

multiple CPMs can be continuously and simultaneously updated in calendar time, constantly learning from new data.

Throughout this section we consider a situation in which we have M previous logistic regression models available to predict a binary outcome Y. These M models have been developed in previous data. For model m, let X_m denoting the design matrix of the covariates; α_m and β_m be the original model intercept and a vector of slopes respectively, and LP stands for linear predictors, so the model is specified by:

$$\text{logit}(P[Y = 1]) = \alpha_m + \beta_m X_m = \alpha_m + LP_m,$$

and $m = 1, \ldots, M$. We wish to update, potentially combine, and apply these models in new data, termed the updating dataset.

2.1 Regression coefficients updating

A simple and widely used strategy is to update the regression coefficients of an existing CPM. This approach can be broadly placed into six ordinal categories based on the extent of modification[4,12]: 1) update the intercept only; 2) update the intercept and adjust the other regression coefficients by a common factor; 3) category 2 plus extra adjustment of a subset of the existing coefficients to a different strength; 4) category 3 plus adding new predictors; 5) re-estimate all of the original regression coefficients; 6) category 5 plus adding new additional predictors. These approaches have been used in various medical applications[23–25 26,27] and were applied to single CPM update.

The first two categories influence the calibration performance of the CPM but not the discrimination. They assume that the relative association between predictors and outcome stays at the same level between the original and new datasets, and the only difference between the two datasets is the observed outcome frequencies. Calibrating the intercept ensures that the observed and expected outcome rate agree on the new dataset. For a given model m this strategy recalculates a new intercept, α_m^{U1}, via fitting a new logistic regression model using the updating dataset while taking the LP_m from the original model as an offset, that is to fix the coefficient for LP_m at unity:

$$\text{logit}(P[Y = 1]) = \alpha_m^{U1} + LP_m.$$

The coefficients for the predictors therefore stay unchanged at β_m for the updated model. Such a method has been proposed in predicting the risk of severe postoperative pain[28].

Method 2 is referred to as "logistic calibration"[29], and estimates an overall correction factor and proportionally adjusts the original coefficients by this factor. It works by fitting a univariate logistic regression using LP_m from the original model m as the covariate. The new predictor-outcome associations are then $\beta_m^{U2} = B_m\beta_m$ with a new intercept to be $\alpha_m^{U2} = A_m$:

$$\text{logit}(P[Y = 1]) = A_m + B_m LP_m.$$

These first two strategies are the simplest and can work effectively when the size of the new dataset is relatively small and the case-mix is similar in the updating and validation sets.

The remaining approaches should be considered if discrimination is of concern, the strength of the association between some predictors and the outcome is thought to be substantially different in the new population, or it may be useful to consider some predictors that were not in the original CPM. Method 3 updates a CPM using method 2 first, then re-estimates a subset of the coefficients that exhibit different strength in the original and new datasets. Some authors[23–25] used objective criteria such as the likelihood ratio test and forward stepwise variable selection to decide which coefficients needed to be adjusted; while others[28,30] used expert knowledge for this decision. Method 4 involves extending the original model by including new risk factors which were not originally in a CPM[30]. These newly added predictors may not have been available when a CPM was first developed, and could lead to further improvement in both calibration and discrimination.

When methods 1 to 4 appear to be inadequate, more extensive revisions can be considered. In Methods 5 and 6, the only way in which the original CPM is used is to select the covariates for inclusion in the model. The historical data is otherwise disregarded. Specifically, Method 5 fits a new CPM based exclusively on the covariates from the existing CPM, i.e. fitting:

$$\text{logit}(P[Y = 1]) = \alpha_m^{U3} + \beta_m^{U3} X.$$

Method 6 additionally allows new predictors to be added. If the original individual-level data are available, these can be combined with the updating dataset to build a new model[12]. Both Methods 5 and 6 are aggressive modelling approaches which give a low or no weighting to the

historical information, are likely to over-fit the data, and the resulting CPMs are less stable. The updated model may fit the local setting perfectly, but may lack external validity.

To overcome the potential problem of over-fitting, applying shrinkage to CPMs estimated by Methods 3 to 6 has been proposed[23], in which the updated coefficients are shrunk either towards zeros or towards the re-calibrated coefficients of Method 2.

2.2 Meta model updating

In the situation where there are multiple historical CPMs available in the literature for the same or similar endpoints and populations (e.g. there are a number of scores[31–33] used for assessing the operative mortality after cardiac surgery in adults), meta-analysis techniques have been proposed to synchronize them into one meta-model, in the presence of an updating dataset[34–36]. Commonly, the original CPMs were derived independently from different populations. The individual-level data used to fit the original CPMs are unlikely to be available, and each of these CPMs may have a distinct set of predictors. By combining these CPMs together, the resulting meta-CPM may have better performance than each individual CPM and be more amenable to generalization to a wider population.

We use the notation above to accompany the main text to explain how a meta model can be created in practice. Assume there are *M* historical CPMs, each with their own design matrix $X_1, \ldots X_M$. The (co)variance among these model coefficients $(\alpha_1, \beta_1), \ldots, (\alpha_M, \beta_M)$ could be obtained; but the original individual-level data may not always be available. A new dataset

with individual-level data *Z=(Y',X')* are available for updating. A meta-model aims to summarize all the information from *M* historical models and the new data *Z* into an overall effect model *CPM*$_T$: $\text{logit}(P[Y=1]) = \alpha_T + \beta_T X$.

A two-stage strategy has been proposed under the scenario when all the historical CPMs and the updating dataset have the same set of predictors[34]. In the first stage, the new dataset *Z* is summarized into estimates of association between outcome and predictors (i.e. to estimate $\text{logit}(P[Y=1]) = \alpha_Z + \beta_Z X$ using data *Z*). This step can be carried out by any method as if a new CPM is first built using the updating dataset. As a result, the new individual data are reduced into model coefficients α_Z and β_Z, and their covariance estimates. In the second stage, traditional meta-analysis techniques are applied to combine the coefficients of $(\alpha_1, \dots, \alpha_M, \alpha_Z)$ and $(\beta_1, \dots, \beta_M, \beta_Z)$ from the new and the historical models together into α_T and β_T. These meta-analysis techniques include: 1) a naïve univariate meta-analysis, which pools estimated effects among various studies via weighted least squares; 2) multivariate meta-analyses with a random effects model considering both within and between studies correlations; and 3) Bayesian inference that use the historical data *CPM*$_1$ … ,*CPM*$_M$ to construct a prior, and use the individual data from updating set *Z* as likelihood; then a new meta-*CPM*$_T$ is formed from the posterior distribution. The intercept of the meta-model is then recalibrated as per the approaches described in Section 2.1. These methods were applied, for example, on a traumatic brain injury and deep venous thrombosis data and it was concluded that the meta-CPM approach improved the discrimination and calibration compared with refitting a new model

using only the updating dataset (ignoring any historical information)[34]. The first two meta-analysis techniques value the historical and the updating datasets equally and produce averaged pooled coefficient effects, therefore it may not predict the target population well as a result. It emphasizes the target updating population more than the historical data. All these strategies assume CPMs share a similar set of predictors, which is not always realistic. Although various remedies had been proposed regarding how to impute the between-risk-factors covariance when they are not available, the robustness of these imputation methods is unknown. It is not clear how to update predictors in the sense of adding or removing predictors under these frameworks.

Debray et al.[35] provide methods for the case where all individual-level data from all the original historical models are available. Their approaches involve re-fitting a meta-model using all the historical individual data, while allowing study-specific intercepts to account for different sources of historical data. That is to create one CPM: $\text{logit}(P[Y=1]) = \alpha_i^U + \beta^U X$ for all source of dataset *i=1….M* with a commonly shared predictor-outcome association β^U for all *i.* This is achieved using a random intercept model, or stratified estimation of the study-specific intercept α_i^U, to account for the heterogeneity caused by different baseline risk from different populations. Various proposals are then made to choose the intercept to use for the new study population. However, this approach only allows the intercept to vary between populations; hence, if the predictor-outcome relationships are highly heterogeneous, the meta-CPM will

perform poorly. As a result, although more information is available, these methods[35] could still be out-performed by a traditional meta-analysis[34] method.

The above updating schemes are limited to the case when all sources of data share the same set of predictors. Model averaging and stacked regression do not have such a constraint[36] (i.e. variables for $X_1,\ldots,X_M$, X' can be all or partially the same, or completely distinct). There are three steps involved in a model averaging meta-model update. The first step involves updating each of the historical CPM via the approaches discussed in Section 2.1. The second step applies Bayesian model averaging on all historical CPMs and obtains weighted average predictions for each individual. The weights are calculated as $w_m = \exp(-0.5BIC_m)/\sum_m \exp(-0.5BIC_m)$; where $BIC_m = -2l_m + k_m \log(N)$, l_m is the log-likelihood, k_m is the number of parameters been updated in the first step (e.g. one parameter for the intercept update), and N is total number of patients in the updating dataset. The third step refits a meta-model using weighted average predictions from the contributing scores as the dependent variable, and using all variables from the original models as independent variables. This approach gives more weight to CPMs which fit the updating dataset better (with higher likelihood) and to those with a less complicated update (penalising those with fewer parameters changed less heavily) in step 1. However, it is not clear how intensive the update should be in step 1; with different strategies, the weights assigned to different historical CPMs would be different and the final meta-model would thus be affected. This model averaging strategy has a tendency to become model selection, assigning a weight of 1 to a single CPM and a weight of 0 to the remaining ones:

because the weight is assigned by an exponential function, a small differences in likelihood or penalizing term could easily inflate a model weight to unity or null.

The stacked regressions meta-model proposal uses the risk score from each CPM as a predictor in a new meta-CPM, which is therefore a logit-linear combination of all pre-existing CPMs[36]. It calculates a weight π_m for each model *m* and updates the coefficients in one go, with the new coefficients for each individual model hence being $\beta_m^{U6} = \beta_m \pi_m$, but overall coefficients being $\beta^{U6} = \sum_m \beta_m \pi_m$. The form of the meta-model using stacked regression is:

$$\text{logit}(P[Y = 1]) = \pi_0 + \sum_m \pi_m LP_m.$$

This strategy uses the updating data less intensively with fewer parameters to be estimated comparing with the model averaging proposal. However, there are clearly multicollinearity issues in the meta-CPM model, and the quoted papers failed to demonstrate whether the stacked regression approach outperformed alternatives using simulation.

In summary, meta-models combine several CPMs into one updated CPM, and have the potential to generalize to a wider population, and to have better performance than the individual CPMs. However, fitting a meta-model may not be practical when there is only a small number of historical CPMs (random effects cannot be estimated, and the weights calculated are unstable). The usual good practice of conducting a meta-analysis, such as selecting well-designed historical CPMs to be combined, should be applied. None of the meta-analysis strategies have discussed adding new predictors. All the meta-model update

techniques described in this section are relatively new applications to the CPM literature. More research is needed to establish principles such as how frequently an update should be carried out, how big the updating dataset should be, and how to conduct model selection.

2.3. Dynamic model (DM) updating

Dynamic updating refers to the continuous updating of one[13] or multiple[37,38] CPMs , as opposed to the previous approaches that are only conducted at fixed time points. As a result, the coefficients for an updated CPM are continuously varying with time. In this section we focus on a Bayesian dynamic logistic regression for a single CPM update ("DM"). For a single model m, let $\theta_m^t = (\alpha_m^t, \beta_m^t)'$ be the vector of parameters for model m at time t. Let Y^t and X^t denote the outcome data and covariate data available up to time t, and y^t and x^t the data from time t only. Let λ be a forgetting parameter. The procedures can be initiated by assuming θ_m^0 is normally distributed with mean estimated at the historical model coefficients (i.e. $\theta_m^0 = (\alpha_m, \beta_m)'$), and covariance matrix Σ^0 estimated (for example) using one-tenth of the updating data (if it is not available from the historical model). Then the *prediction equation* is:

$$p(\theta_m^t | Y^{t-1}, X^{t-1}) = N(\theta_m^{t-1}, R^t); \quad R^t = \lambda^{-1}\Sigma^{t-1}.$$

The *updating equation* is proportional to the product of a Bernoulli density (Likelihood) and the prediction equation (Prior) so that the whole procedure has a Bayesian interpretation:

$$p(\theta_m^t | Y^t, X^t) \propto p(y^t | \theta_m^t) p(\theta_m^t | Y^{t-1}, X^{t-1}) \propto Likelihood \times Prior$$

The estimate of θ_m^t is chosen to maximise this expression; the expression cannot be written in closed form, so is approximated using a Normal distribution.

The forgetting parameter λ is embedded in the *prediction equation* which controls the variance of the prior distribution in the *updating equation*: with a small λ, it is equivalent to have a less informative prior that the updated model will rely less on the historical information (through a flatter prediction equation).

The estimating procedure involves recursively applying the prediction and the updating steps.

The DM updating can also be applied to several CPMs simultaneously, resulting in dynamic model averaging (DMA)[37,38]. This assumes that there are multiple historical CPMs acting together at all times; at a given time, some are more predictive than others, and how well a CPM predicts may alter over time. A physical example might refer to the existence of several latent sub-populations, and the proportions of these sub-populations in a target population may vary over time. Each of these sub-populations can be predicted by a specific CPM model. Therefore, DMA can be used to predict an optimal weighted average for the whole population at any time point.

DMA has been applied to a continuous outcome in an engineering cold rolling mill example[37] and to a binary outcome medical example of paediatric laparoscopic appendectomies[38]. In the latter medical example, the purpose of the study was for inference rather than future prediction, and we are not aware of any DMA application on updating CPMs. However, a single

DM update has been applied in cardiac surgery data[10]. In general, DM can continuously adapt to the changes of the underlying process, without paying a high price for model uncertainty in a big model space, and is relatively insensitive to the choice of forgetting parameters except when responding to abrupt changes[37]. The forgetting parameters can be selected using an auto-tuning procedure. Although this procedure can be conducted at each time point, for each parameter, and on a continuous scale, this has a high computational load. Therefore, a simplified discrete proposal on the choice of forgetting parameters has been suggested for computational feasibility[38]. DM incorporates historical data, which is likely to provide a smooth and stable update to the coefficients[13]. Most of the applications so far are explanatory in nature to study the relationship between outcome and predictors, less so for the prediction purpose.

Comparing DMA with meta-CPM using Bayesian model averaging (BMA) approach[36], the latter case considered a fixed set of 'true' models while the former method allowed the multiple 'true' models to vary over time. If any static strategy were applied repeatedly over time, this would itself become a dynamic approach[39,40].

To summarize, DMs are more adaptive than the single static model solutions described in Sections 3.1 and 3.2. Although a static model can be updated repeatedly, these methods would only be implemented when there are fair amounts of new data available, and consequently they should not be conducted too often. The risk factors of a CPM are likely to evolve slowly over time; DM responds to this in a smooth way. DMA has the ability to consider

a huge model space, in that case it might be viewed as an *automatic model selection* process by implementing this approach on all 2^k possible models (with k covariates), and one can identify the most active models using posterior model probabilities. However the computing load might not be trivial. Moreover, it is not clear how a dynamic model should be validated, and there may be issues with clinical acceptability if outputs, and hence recommended clinical decisions may be changing from one day to the next.

We are now turning towards evaluating the different proposals. To do so, we will use data from the National Adult Cardiac Surgery Audit (NACSA) registry, which is described in more detail in the next section.

3. National Adult Cardiac Surgery Audit (NACSA) Registry Data

Members of the Society for Cardiothoracic Surgery in Great Britain and Ireland (SCTS) submit clinical data on adult cardiac surgery operations to the National Adult Cardiac Surgery Audit (NACSA) registry—one of six national clinical audit databases managed by the National Institute of Cardiovascular Outcomes Research (University College London), covering all National Health Service trusts and some private and Irish hospitals. The SCTS have been recording data in some form since 1977, and have published risk-adjusted mortality outcomes on the individual surgeon level since 2005[11].

Historically, the commonly used risk score for auditing and decision-making in cardiac surgery in the UK was the logistic EuroSCORE[41] (ES), which was published to replace the additive EuroSCORE risk score[42]. This is a logistic regression model that produces a predicted probability of mortality for patients due to any cause following a cardiac operation, based on risk factors available before a procedure is carried out. ES is based on data collected in 1995 (although the logistic model was not published until 2003), and over time it has become poorly calibrated[43]. More recently this has been replaced with EuroSCORE II[44,45] (ES2), which is based on data collected in 2010. Whilst structurally similar to ES, ES2 updates the definition of some original risk factors. For example, ES did not differentiate risk between a patient having an isolated mitral valve repair operation and a patient have quadruple coronary artery bypass surgery, mitral valve repair and aortic valve replacement, whereas ES2 introduces a 'weight of intervention' variable. Furthermore, ES2 incorporates some new factors, and removes others.

The complete NACSA registry was downloaded and pre-processed using cleaning rules developed in collaboration with cardiac surgeons[46]. These rules were employed to harmonise transcriptional discrepancies, map data between different database versions, remove clinically implausible values, and remove paediatric, duplicate and non-cardiac surgery records. Following this all records between 1st April 2007 and 31st March 2012 were retained. Figure 1 is a flow chart showing various subsets of NACSA data. There are 182,492 records during this five year period. As the focus of this paper is illustration of methodology rather than clinical use, we restricted our subjects to those who received coronary artery bypass graft surgery (CABG)

either isolated or with other concomitant cardiothoracic surgery. However, we note that the EuroSCORE models were developed for prediction in the population of all cardiac surgery patients, not a specific procedural subgroup. We conducted our analyses on a single imputed dataset derived using the chained equations technique[47,48]. There were 127,946 patients after imputation. Complete-case analysis was also carried out as a sensitivity analysis (114,345 complete cases) and results for this are reported in the *Supplemental Material*. Construction of the datasets is reported in Figure 1 and Table 1.

The NACSA dataset was chronologically ordered according to the date of surgery, then split in a 2:1 ratio with the first two-thirds of the data (n=85,297) used for updating; and the remaining third (N= 42,649) used for validation. As the two datasets are from difference time periods, this constitutes transportability rather than internal validation[49].

4. Analysis and Results

4.1 Comparative evaluation setup

In order to demonstrate some of the methodologies described in Section 2 and highlight differences between them, we chose ES[41] and ES2[44] to be updated using the NACSA updating dataset. Once the update was considered to be satisfactory, the updated models were tested on the NACSA validation dataset.

There is evidence that both ES and ES2 are miscalibrated on contemporary NACSA data, due to differences in characteristics between NACSA and the original datasets on which the two scores were derived[43,45]. Moreover, the predicted (risk-adjusted) mortality has increased over time in the dataset from 6.09% and 2.93% (2007/08) to 6.51% and 3.26% (2011/12) for ES and ES2 respectively.

The crude mortality in the updating dataset is 2.67% in contrast to 4.8% and 3.9% for the data on which ES and ES2 are based on respectively. Applying both scores directly to the updating dataset, they showed good discrimination (AUC for ES: 0.817 (95% CI: 0.809, 0.826); for ES2: 0.831 (95% CI: 0.823, 0.839)), but over-predicted the mean mortality to be 6.16% using ES and 2.94% using ES2. The miscalibration of ES is also reflected in a logistic regression of the outcome on the predicted log-odds yielding slopes (0.97 (SE=0.009)) and intercepts (-0.99 (SE=0.022)) different from 1 and 0 respectively. These results highlight the need for exploring revised models for use in the NACSA data.

Table 1 Imputed data analysis: Summary of the original, updating, and validation datasets .

	ES original data	**ES2 original data**	**NACSA updating**	**NACSA validation**
Data collection period	Sep-Nov 1995	3rd May-25th Jul 2010	1st April 2007-26th May 2010	26th May 2010-31st Mar 2012
Sample size	19,030	22,381	85,297	42,649
Setting	128 surgical	154 hospitals,	43 hospitals, 1	43 hospitals, 1

	centres, 8 countries*	43 countries*	country(UK)	country(UK)
Mortality %	4.8%	3.9%	2.67%	2.69%

*Data from UK were included in both ES and ES2.

In the remainder of this section we illustrate six of the updating strategies:

- ***Strategy I: Intercept update (Category 1 in Section 2.1).***
- ***Strategy II: Logistic calibration (Category 2 in Section 2.1).***
- ***Strategy III: Model refit (Category 5 in Section 2.1).***
- ***Strategy IV: Dynamic updating using Bayesian dynamic logistic model for single CPM (DM).*** For the current analysis λ is treated as a time-invariant scalar, fixed at 0.99 for the main analysis with sensitivity analysis conducted for other values of λ (see Section 4.3). Updates were made on a monthly basis. We use the updated CPM on 5th May 2010 (the final date of the updating data) for validation purposes.
- ***Strategy V Meta-model with model averaging update.*** For this analysis, almost all the weight was placed on ES2 hence this strategy resembles a model selection procedure (the weight for ES is 5.2×10^{-71}). We therefore also evaluated a pragmatic approach of assigning equal weights to both CPMs.

Most of the ES2 predictors are not new to ES but change their definitions from ES, while 9 predictors share exactly the same definitions in ES and ES2. For those re-defined variables, we adopt their newer definitions.

- ***Strategy VI: Meta-model with stack regression update.*** For the initial recalibration stage, we used intercept updating only.

All analyses were conducted using R[50] (version 3.0.1). R Packages *pROC*[51] and *dma*[52] were used to calculate AUC and for DM modelling respectively. The (Cox) calibration[21] intercept and slope were estimated using univariate logistic regression. Calibration plots[22] were produced using lowess smoother. R codes are available on request from the authors.

4.2 Validation results

Table 2 summarises the calibration and discrimination performance of an updated CPM using the validation dataset based on the six selected strategies.

When the original ES was applied directly to the validation dataset, it showed good discrimination (AUC=0.819) but it systematically over-estimated the mortality (calibration intercept=-1.06, slope=0.97). All updating strategies greatly improved the calibration over the original model and slightly improved discrimination. Adjusting the model intercept (Strategy I) contributed the most toward calibration improvement with the calibration intercept changing from -1.06 to -0.07. All Strategies II to IV showed further improvement of calibration and

discrimination over Strategy I which suggested the strength of predictors and outcomes association is different in the original ES and the NACSA datasets.

The unmodified ES2 showed good discrimination (AUC=0.828) but slightly overestimated the mortality (calibration intercept of -0.20) in the validation dataset. With simple intercept adjustment (Strategy I) the updated CPM showed good fit to the validation data, with the calibration slope and intercept becoming very closed to1 and 0 respectively(although calibration intercept is still statistically significantly different from zero). Here, the updated models using Strategies II to IV made little or no improvement to model performance, compared with Strategy I.

The meta-model Strategy V using pragmatic equal weights approach performed better than each of the individual original CPMs, mainly as a result of both intercepts of the original CPMs being updated before the two models were combined. With one extra predictor than the original ES2, this meta-model had better calibration than the original ES2. Strategy VI of stacked regression modelling performed better than ES and ES2 although this improvement is slight; as expected it assigned more weight to ES2 ($\pi_1 = 0.90$) than ES ($\pi_2 = 0.18$), but the difference is far less extreme than strategy V.

As an alternative to Cox recalibration, the calibration plots of the binary observed outcomes regressed on the predicted probabilities from an updated CPM using non-parametric *loess* smoother were also examined (Figure 2). The results showed similar patterns to the results

reported above, that not a single method out-performed others. The lack of calibration was mainly demonstrated in the high predicted probabilities range.

The complete-case analysis is reported in *Supplemental Material*. For all 6 updating strategies chosen, they showed similar discrimination and calibration to the imputed-data analyses.

Table 2. Imputed data analysis: Performance of updated CPMs on the validation dataset.

	Discrimination		**Cox recalibration**	
Strategy	**AUC (95% CI)**		**Intercept (SE)**	**Slope (SE)**
Original EURO1 (ES)	0.8194 (0.8079-0.8309)		**-1.06092 (0.031678)***	**0.97285 (0.013164)***
I	0.8194 (0.8079-0.8309)		**-0.06656 (.031679)***	0.98535 (0.009733)
II	0.8194 (0.8079-0.8309)		-0.05877(0.031291)	1.00109 (0.009745)
III	0.8293(0.818-0.8405)		-0.03461(0.031345)	1.00293 (0.009856)
IV (λ =0.99)	0.8293(0.8181-0.8405)		-0.05159(0.031299)	1.00534 (0.009892)
Original EURO2 (ES2)	0.8279(0.8164-0.8394)		**-0.20358(0.031269)***	1.00855 (0.010277)
I	0.8279(0.8164-0.8394)		**-0.09525(0.031269)***	1.00803 (0.009952)
II	0.8279(0.8164-0.8394)		**-0.10722 (0.031611)***	0.99432 (0.009964)
III	0.8344(0.8231-0.8457)		**-0.10882(0.031611)***	0.99588 (0.010181)
IV(λ=0.99)	0.8345(0.8232-0.8458)		**-0.10719(0.031557)***	0.99874 (0.010170)
Meta-model				
V	0.8289(0.8176-0.8402)		**-0.09071(0.031405)***	1.00267(0.009944)
VI	0.8297(0.8183-0.841)		**-0.09920(0.031595)***	0.99588(0.009962)

∗ Values of intercept and slope in bold are significantly different from 0 or 1 respectively, using a p-value cutoff of 0.05.

4.3 Sensitivity analysis on DM parameter λ

Sensitivity analysis was conducted on the forgetting parameter λ, for a range of values between 0.5 and 1. For both ES and ES2, the best AUC and calibration were achieved when λ was close to 0.99. Despite λ being close to 1, there is still a significant amount of forgetting, because the forgetting operates in a compound manner, and there are 38 time points in the training dataset. A smaller forgetting parameter is equivalent to using a less informative prior in the updating equation stage, so that the data from the far past become less influential on the current dynamic model estimates. Using a smaller λ, the estimation relies on a smaller recent dataset, which made the validation results worse than with a larger λ.

We did not adopt dynamic model averaging here, but a single CPM dynamic model approach. We could input both ES and ES2 into a single DMA framework as two potential models and allow them to be active at different time and with different rates. However, due to the similar nature of these two models, they do not capture different dimensions to predict the mortality.

5. Discussion

In this paper we have reviewed a variety of approaches that can revise CPMs for a new population and maintain performance over time, and we have contrasted some of the key methods using a typical example of mortality surveillance around cardiac surgery.

In the cardiac surgery example there was general agreement that an older risk score (logistic EuroScore; ES) needed to be updated while this was less evident for a more recent score (EuroScore II; ES2). Following Hickey et al[13], the need for updating of ES is likely to be

attributed to a change in case-mix and various characteristics in the data that have changed over the years. In particular, the case-mix adjusted mortality rate had decreased substantially.

In comparing the various updating approaches, we did not find any single method that outperformed the others, the differences are more nuanced. For situations where only small changes between the original and updating datasets occur, simple re-calibration methods, such as intercept updates, are sufficient as seen on the example of ES2. More involved methods are useful when larger changes are evident, as seen in the case of ES. It is also noteworthy here that we used a large updating dataset, which will support more complex updating strategies. Meta-model approaches are a new area and more research is needed before these approaches can be recommended, particularly around dealing with the high multicollinearity between the risk scores. Dynamic modelling is a promising area allowing continuous updating, which is particularly relevant given the trend towards instant data capture and regular uploads to a central database. One distinguishing feature of the various approaches is the quantity of information on previous data and models that are required to utilize a given approach, and how this past information is integrated into the new model.

The validation we have provided is officially testing transportability rather than internal validation. However, the similarity of the time periods means this could be viewed as a split-sample validation. From this perspective, such a method is known to be outperformed by bootstrapping base validation[53]. However, at the level of 38.3 events per variable in validation set (and 75.8 in updating set) the validation performance of these two methods are similar [53].

Other than the aforementioned methods, there are some pragmatic updating strategies which make ad-hoc adjustments to the predicted risks. For example, a calibration factor is published every quarter for The Society of Thoracic Surgeons National Adult Cardiac Surgery risk[54]. In order to consider clustering effects, at e.g. hospital level, methods such as random effects models or generalized estimating equations (GEEs) have been suggested[31]. However, such models may be difficult for clinical users to understand.

Sophisticated updating based on a very small dataset should be applied with caution, because the methods discussed here put comparatively large weight on the updating data and hence the CPM would be prone to peculiarities of the updating dataset, and hence over-fitting. One potential remedy for this issue is to shrink the coefficient estimates towards the original CPM according to the relative sizes of the original and updating datasets.

Although more complex approaches, such as DM or DMA, are attractive from a statistical standpoint, it is unclear how practical such approaches are. They will be more difficult to explain to non-statisticians, more difficult to validate, and implementation requires a continuous data stream.

In conclusion, there are a wide range of updating methods available for applying CPMs in new populations and maintaining their performance over time – ranging from very simple calibration adjustments, to more involved approaches involving dynamic modelling or combining multiple CPMs. Whenever there are existing CPMs that have the potential to apply to the clinical domain or population of interest, their performance, and potential for updating

or revision, should be explored before considering the development of a new CPM from scratch. This will help to avoid the perplexing quantity of CPMs operating in similar or identical contexts. Although this article only focus on statistical aspect, a CPM should not only be judged by its quantitative performance; equally important is its clinical and face validity.

Acknowledgements

This work is independent research arising in part from Dr. Jaki's Career Development Fellowship (NIHR-CDF-2010-03-32) supported by the National Institute for Health Research (NIHR), NIHR Research Methods Opportunity Funding Scheme (RMOFS 2012-09), and the Medical Research Council funded Health e-Research Centre (MR/K006665/1).

The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council, or the Department of Health.

Access to the National Adult Cardiac Surgery Audit registry data was approved by the National Institute of Cardiovascular Outcomes Research (NICOR), University College London (UCL), research board. Details of the data sharing agreements can be found at http://www.ucl.ac.uk/nicor/access/application.

The authors declare that there is no conflict of interest.

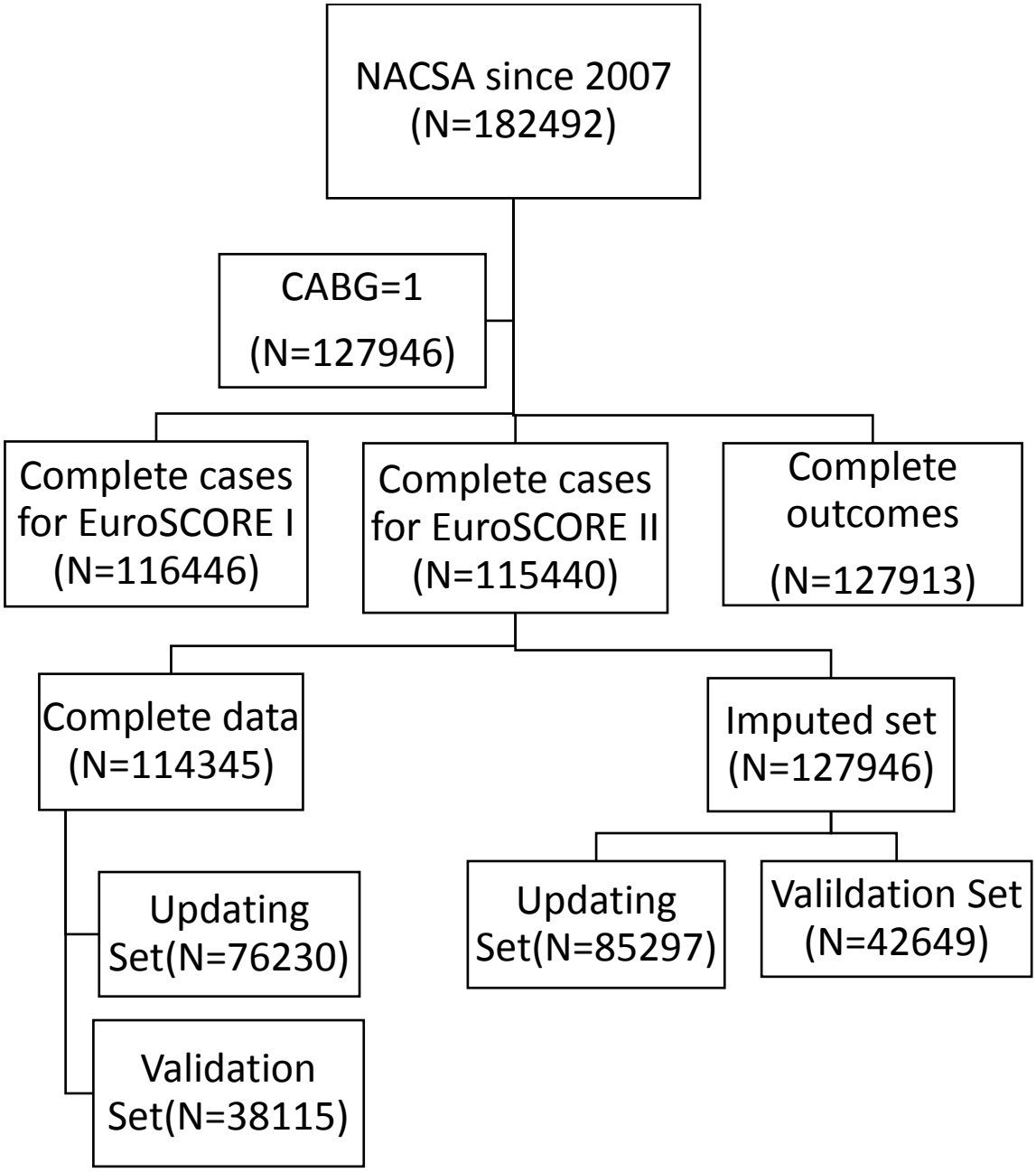

Figure 1: A flow chart of NACSA registry data structure. CABG=1 identified patients who underwent coronary artery bypass graft surgery.

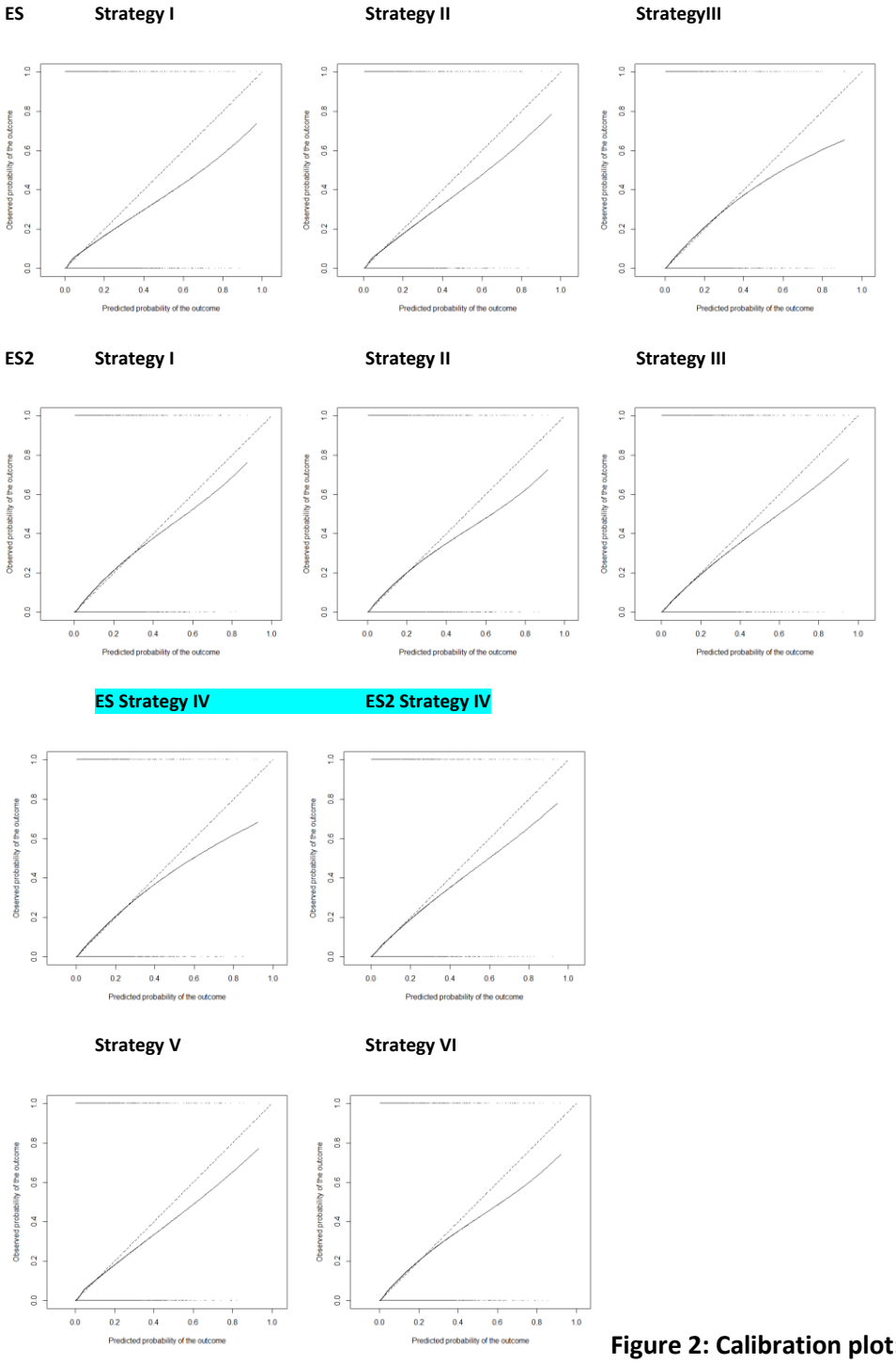

Figure 2: Calibration plot: Diagonal line has intercept=0 and slope=1. The binary outcomes were displayed as dots along y=0/1 lines.

References

1. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer; 2008.

2. Hingorani AD, van Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ Br Med J.* 2013;346.

3. Beattie P, Nelson R, others. Clinical prediction rules: What are they and what do they tell us? *Aust J Physiother.* 2006;52(3):157.

4. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21(1):128.

5. Lovegrove J, Valencia O, Treasure T, Sherlaw-Johnson C, Gallivan S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet.* 1997 Oct;350(9085):1128–30.

6. Rogers CA, Reeves BC, Caputo M, Ganesh JS, Bonser RS, Angelini GD. Control chart methods for monitoring cardiac surgical performance and their interpretation. *J Thorac Cardiovasc Surg.* 2004 Dec;128(6):811–9.

7. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ.* 2013;346.

8. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis research strategy (PROGRESS) 2: Prognostic factor research. *PLoS Med.* 2013;10(2):e1001380.

9. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med.* 2013;10(2):e1001381.

10. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med.* 2015 Jan 6;162(1):55.

11. Bridgewater, B., B. Keogh, R. Kinsman and PW. The Society for Cardiothoracic Surgery in Great Britain & Ireland, 6th national adult cardiac surgical database report; demonstrating quality, 2008.

12. Janssen K, Moons K, Kalkman C, Grobbee D, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol.* 2008;61(1):76–86.

13. Hickey GL, Grant SW, Caiado C, Kendall S, Dunning J, Poullis M, et al. Dynamic Prediction Modeling Approaches for Cardiac Surgery. *Circ Cardiovasc Qual Outcomes.* 2013;6(6):649–658.

14. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *Bmj.* 2009;338:1432–1435.

15. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett.* 2006 Jun;27(8):861–874.

16. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982 Apr;143(1):29–36.

17. Gengsheng Qin, Hotilovac L. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Stat Methods Med Res.* 2008 Apr;17(2):207–21.

18. Corinna Cortes MM. Confidence intervals for the area under the ROC curve. *Adv Neural Inf Process Syst.* 2005;17:305–312.

19. DeLong ER, DeLong DM C-PD. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach on JSTOR. *Biometrics.* 1988;44:837–844.

20. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Commun Stat Methods.* 1980;9(10):1043–1069.

21. Cox DR. Two futher applications of a model for binary regression. *Biometrika.* 1958;45:562–565.

22. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med.* 2014;33(3):517–535.

23. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med.* 2004 Aug 30;23(16):2567–86.

24. Janssen KJM, Kalkman CJ, Grobbee DE, Bonsel GJ, Moons KGM, Vergouwe Y. The risk of severe postoperative pain: modification and validation of a clinical prediction rule. *Anesth Analg.* 2008 Oct;107(4):1330–9.

25. Van Hoorde K, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW, Van Calster B. Simple dichotomous updating methods improved the validity of polytomous prediction models. *J Clin Epidemiol.* 2013 Oct;66(10):1158–65.

26. Van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat Med.* 1995 Oct 30;14(18):1999–2008.

27. Van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Stat Med.* 2000 Dec 30;19(24):3401–15.

28. Janssen KJM, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KGM. A simple method to adjust clinical prediction models to local circumstances. *Can J Anaesth.* 2009 Mar;56(3):194–201.

29. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* 1996 Feb 28;15(4):361–87.

30. Kappen TH, Vergouwe Y, van Klei WA, van Wolfswinkel L, Kalkman CJ, Moons KGM. Adaptation of clinical prediction models for application in local settings. *Med Decis Making.* Jan;32(3):E1–10.

31. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1--coronary artery bypass grafting surgery. *Ann Thorac Surg.* 2009 Jul;88(1 Suppl):S2–22.

32. O'Brien SM, Shahian DM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2--isolated valve surgery. *Ann Thorac Surg.* 2009 Jul;88(1 Suppl):S23–42.

33. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3--valve plus coronary artery bypass grafting surgery. *Ann Thorac Surg.* 2009 Jul;88(1 Suppl):S43–62.

34. Debray TPA, Koffijberg H, Vergouwe Y, Moons KGM, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. *Stat Med.* 2012 Oct 15;31(23):2697–712.

35. Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med.* 2013 Aug 15;32(18):3158–80.

36. Debray TPA, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW, Moons KGM. Meta-analysis and aggregation of multiple published prediction models. *Stat Med.* 2014 Jun 30;33(14):2341–62.

37. Raftery AE, Kárný M, Ettler P. Online Prediction Under Model Uncertainty via Dynamic Model Averaging: Application to a Cold Rolling Mill. *Technometrics.* 2010 Feb;52(1):52–66.

38. McCormick TH, Raftery AE, Madigan D, Burd RS. Dynamic logistic regression and dynamic model averaging for binary classification. *Biometrics.* 2012 Mar;68(1):23–30.

39. Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Mon. Weather Rev.* 2005;133:1155–1174.

40. Strobl AN, Vickers AJ, Calster B van, Steyerberg E, Leach RJ, Thompson IM, et al. Improving Patient Prostate Cancer Risk Assessment: Moving From Static, Globally-Applied to Dynamic, Practice-Specific Cancer Risk Calculators. *J Biomed Inform.* 2015 May;

41. Roques F. The logistic EuroSCORE. *Eur Heart J.* 2003 May;24(9):882.

42. Nashef SAM, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardio-Thoracic Surg.* 1999;16(1):9–13.

43. Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K, et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur J Cardio-Thoracic Surg.* 2013;43(6):1146–52.

44. Nashef S a M, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. EuroSCORE II. *Eur J Cardiothorac Surg.* 2012 Apr;41(4):734–44; discussion 744–5.

45. Grant SW, Hickey GL, Dimarakis I, Trivedi U, Bryan A, Treasure T, et al. How does EuroSCORE II perform in UK cardiac surgery; an analysis of 23 740 patients from the Society for Cardiothoracic Surgery in Great Britain and Ireland National Database. *Heart.* 2012 Nov;98(21):1568–72.

46. Hickey GL, Grant SW, Cosgriff R, Dimarakis I, Pagano D, Kappetein AP, et al. Clinical registries: governance, management, analysis and applications. *Eur J Cardiothorac Surg.* 2013 Oct;44(4):605–14.

47. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* 2011 Dec 21;

48. Little, Roderick J, Rubin DBR. Statistical Analysis with Missing Data, 2nd Edition. Second. Wiley; 2002.

49. Justice AC. Assessing the Generalizability of Prognostic Information. *Ann Intern Med.* 1999 Mar 16;130(6):515.

50. {R Core Team}/R Foundation for Statistical Computing. R: A Language and Environment for Statistical Computing [Internet]. 2014;Available from: http://www.r-project.org/

51. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.* 2011 Jan;12:77.

52. McCormick TH, Raftery AE, Madigan D. dma: dynamic model averaging. [Internet]. 2012;Available from: http//cran.r-project.org/package=dma

53. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res.* 2014 Nov 19;

54. Jin R, Furnary AP, Fine SC, Blackstone EH, Grunkemeier GL. Using Society of Thoracic Surgeons risk models for risk-adjusting cardiac surgery results. *Ann Thorac Surg.* 2010 Mar;89(3):677–82.

Supplemental Material: Complete-case analysis

Table S1. Summary of the original, updating, and validation datasets.

	ES original data	**ES2 original data**	**NACSA updating**	**NACSA validation**
Data collection period	Sep-Nov 1995	3rd May-25th Jul 2010	1st April 2007-5th May 2010	5th May 2010-31st Mar 2012
Sample size	19,030	22,381	76,230	38,115
Setting	128 surgical centres, 8 countries*	154 hospitals, 43 countries*	43 hospitals, 1 country(UK)	43 hospitals, 1 country(UK)
Mortality %	4.8%	3.9%	2.6%	2.5%

*Data from UK were included in both ES and ES2.

Table S2. Performance of updated CPMs on the validation dataset.

	Discrimination		**Cox recalibration**	
Strategy	**AUC (95% CI)**		**Intercept (SE)**	**Slope (SE)**
Original EURO1 (ES)	0.8155 (0.8027-0.8283)		**-1.13477(0.034525)***	**0.96628(0.014281)***
I	0.8155 (0.8027-0.8283)		**-0.13324(0.034525)***	0.98182(0.010569)
II	0.8155 (0.8027-0.8283)		**-0.12329(0.034204)***	0.99448(0.010579)
III	0.8268(0.8143-0.8393)		**-0.10687(0.034273)***	0.99667(0.010747)
IV (λ =0.99)	0.8265 (0.814-0.839)		**-0.13333(0.034228)***	0.99895(0.010815)
Original EURO2 (ES2)	0.8245 (0.8118-0.8372)		**-0.27264(0.034068)***	1.00694(0.011196)
I	0.8245 (0.8118-0.8372)		**-0.14656(0.034068)***	1.00644(0.010787)
II	0.8245 (0.8118-0.8372)		**-0.16590(0.034533)***	0.98860(0.010804)
III	0.8319 (0.8194-0.8444)		**-0.17494(0.034520)***	0.99037(0.011078)
IV(λ=0.99)	0.8318 (0.8193-0.8443)		**-0.19441(0.034472)***	0.99276 (0.011134)

Meta-model				
V	0.8247 (0.8121-0.8373)		**-0.14833(0.034225)***	1.00010 (0.010782)
VI	0.8261 (0.8135-0.8387)		**-0.15935(0.034520)***	0.98983 (0.010803)

∗ Values of intercept and slope in bold are significantly different from 0 or 1 respectively, using a p-value cutoff of 0.05.