Martin et al. BMC Medical Research Methodology (2017) 17:1
DOI 10.1186/s12874-016-0277-1

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Clinical prediction in defined populations: a simulation study investigating when and how to aggregate existing models

Glen P. Martin[1*], Mamas A. Mamas[1,2], Niels Peek[1,3], Iain Buchan[1,3] and Matthew Sperrin[1]

Abstract

Background: Clinical prediction models (CPMs) are increasingly deployed to support healthcare decisions but they are derived inconsistently, in part due to limited data. An emerging alternative is to aggregate existing CPMs developed for similar settings and outcomes. This simulation study aimed to investigate the impact of between-population-heterogeneity and sample size on aggregating existing CPMs in a defined population, compared with developing a model de novo.

Methods: Simulations were designed to mimic a scenario in which multiple CPMs for a binary outcome had been derived in distinct, heterogeneous populations, with potentially different predictors available in each. We then generated a new 'local' population and compared the performance of CPMs developed for this population by aggregation, using stacked regression, principal component analysis or partial least squares, with redevelopment from scratch using backwards selection and penalised regression.

Results: While redevelopment approaches resulted in models that were miscalibrated for local datasets of less than 500 observations, model aggregation methods were well calibrated across all simulation scenarios. When the size of local data was less than 1000 observations and between-population-heterogeneity was small, aggregating existing CPMs gave better discrimination and had the lowest mean square error in the predicted risks compared with deriving a new model. Conversely, given greater than 1000 observations and significant between-population-heterogeneity, then redevelopment outperformed the aggregation approaches. In all other scenarios, both aggregation and de novo derivation resulted in similar predictive performance.

Conclusion: This study demonstrates a pragmatic approach to contextualising CPMs to defined populations. When aiming to develop models in defined populations, modellers should consider existing CPMs, with aggregation approaches being a suitable modelling strategy particularly with sparse data on the local population.

Keywords: Clinical prediction models, Model aggregation, Validation, Computer simulation, Contextual heterogeneity

Background

Clinical prediction models (CPMs), which compute the risk of an outcome for a given set of patient characteristics, are fundamental to clinical decision support systems. For instance, practical uses of CPMs include facilitating discussions about the risks associated with a proposed treatment strategy, assisting audit analyses and benchmarking post-procedural outcomes. Consequently, there is growing interest in developing CPMs to support local healthcare decisions [1, 2]. Although there might be existing models derived for similar outcomes and populations, it is vital they are appropriately updated, validated and transferred between different contexts of use. Baseline risk and predictor effects may differ across populations, which can cause model performance to decrease when transferring an existing CPM to the local population [3–6]. This between-population-heterogeneity frequently leads to researchers rejecting existing models

* Correspondence: glen.martin@manchester.ac.uk
[1]Health e-Research Centre, University of Manchester, Vaughan House, Portsmouth Street, M13 9GB Manchester, UK
Full list of author information is available at the end of the article

and developing new ones [5, 7–10]. However, this framework is undesirable because the dataset used to derive the new CPM is often smaller than previous derivation datasets and can lead to multiple models for the same outcome. For instance, over 60 previously published models predict breast cancer [11], which is perplexing and unhelpful to end-users.

As a motivating example, consider a user wishing to predict short-term mortality post cardiac-surgery. There are several existing CPMs available, including the Logistic EuroSCORE (LES), the EuroSCORE II (ESII), the STS score and the German Aortic Valve Model (German AV) [12–16]. These models share some common predictors, for example gender, arterial disease outside the heart and recent heart attack, but some predictors appear only in a subset of the CPMs. For instance, diabetes is only incorporated into the ESII and STS models, while atrial fibrillation is only in the STS and German AV models. Moreover, definitions and coding of some predictors could differ: examples include left ventricular ejection fraction and age.

While differences in variable definitions between existing CPMs add complexity, the prior information encapsulated by each model can be exploited. A generalizable existing CPM could serve as an informative prior for a new population; for example, by transferring information regarding likely covariate-outcome relations, as in stacked regression [9, 17]. However, there has been limited investigation into the impact of sample size and between-population-heterogeneity on the performance of model aggregation versus deriving a new CPM.

This simulation study considers a situation in which there is a new (local) population, with associated data, and interest lies in developing a CPM for it. The modeller must make a choice between utilising existing CPMs that have been developed in different populations, developing a new model and disregarding existing ones, or some mixture of the two [18]. We hypothesised that the modelling strategy that optimised performance would depend on: 1) the degree of variation in risk across multiple populations (between-population-heterogeneity); and 2) the quantity of data available in the local population, relative to the size of previous derivation datasets.

Methods

Throughout this study, all CPMs will be assumed to be logistic regression models, although the techniques apply to other types of prediction model, such as those for time-to-event outcomes. Stacked regression (SR) [9, 17], principal component analysis (PCA) [19, 20] and partial least squares (PLS) are three possible methods that simultaneously aggregate and calibrate existing models to a new population. We describe SR and PCA here, with PLS described in Additional file 1. This study compares the three aforementioned aggregate approaches with deriving a new model; possible techniques of redevelopment are also outlined in this section.

Model aggregation: stacked regression

Consider a collection of M existing logistic regression CPMs, which all aim to predict the same binary outcome but were derived in different populations, $j = 1, \ldots, M$. For a set of observations $i = 1, \ldots, n_j$ from population j, let $\boldsymbol{X}_j$ denote the $n_j \times P$ matrix of predictors that are potentially associated with the outcome, $\boldsymbol{Y}_j$. Here, P represents the number of predictors available across all populations; a predictor that is not present in a given CPM simply has coefficient zero. Then, for $i = 1, \ldots, n_j$, the linear predictor (LP) from the j^{th} existing CPM, $\text{LP}_{i,j}$, is given by

$$\text{LP}_{i,j} = \beta_{0,j} + \sum_{p=1}^{P} \beta_{p,j} x_{i,p}$$

with intercept $\beta_{0,j}$ and coefficients $\beta_{1,j}, \ldots, \beta_{Pj}$, at least one of which is non-zero.

Suppose we then have a new local population, $j = M + 1$. Stacked regression aims to weight the M linear predictors (calculated for each observation in the new local population) to maximise the logistic regression likelihood. Specifically, SR assumes that for $i = 1, \ldots, n_{M+1}$, $Y_{i,M+1} \sim \text{Bernoulli}(\pi_{\text{i,M}+1})$ where $\pi_{i,M+1} = P(Y_{i,M+1} = 1)$ with

$$\log\left(\frac{\pi_{i,M+1}}{1-\pi_{i,M+1}}\right) = \hat{\gamma}_0 + \sum_{j=1}^{M} \hat{\gamma}_j \text{LP}_{i,j}$$

under the constraint that $\hat{\gamma}_1, \ldots, \hat{\gamma}_M \geq 0$ to account for collinearity between the existing CPMs. Here, $\text{LP}_{i,j}$ denotes the linear predictor from the j^{th} existing CPM calculated for observation $i \in [1, n_{M+1}]$ in the new local population. Thus, we have

$$\log\left(\frac{\pi_{i,M+1}}{1-\pi_{i,M+1}}\right) = \hat{\gamma}_0 + \sum_{j=1}^{M} \hat{\gamma}_j \beta_{0,j} + \sum_{p=1}^{P} \sum_{j=1}^{M} \hat{\gamma}_j \beta_{p,j} x_{i,p},$$

which can be used to calculate subsequent risk predictions for a new observation. The hat accent above parameters indicates those that are estimated from the local population data. Specifically, SR estimates $\hat{\gamma}_j$ but not $\beta_{p,j}$, which are taken as fixed values from the published existing CPM.

Model aggregation: Principal Components Analysis (PCA) regression

Let $\mathbf{LP}$ denote the $n_{M+1} \times M$ matrix, with $(i,j)^{\text{th}}$ element being the linear predictor for the j^{th} existing CPM calculated for observations $i = 1, \ldots, n_{M+1}$ in the local population. The singular value decomposition of $\mathbf{LP}$ gives an $M \times M$ rotation matrix, $\boldsymbol{\nu}$. Multiplying $\mathbf{LP}$ by $\boldsymbol{\nu}$, gives the $n_{M+1} \times M$ matrix of principal components, $\mathbf{Z}$. PCA

regression again assumes that $Y_{i,M+1} \sim \text{Bernoulli}(\pi_{i,M+1})$ for $i = 1, \ldots, n_{M+1}$ with

$$\log\left(\frac{\pi_{i,M+1}}{1-\pi_{i,M+1}}\right) = \hat{\theta}_0 + \sum_{j=1}^{M}\hat{\theta}_j Z_{i,j}.$$

Unlike in SR, no restrictions are placed on the parameters $\hat{\theta}_j$ since, by definition, the principal components, $\mathbf{Z}$, are uncorrelated. One can obtain predictions for a future observation by converting the above aggregate model onto the scale of the original risk factors,

$$\log\left(\frac{\pi_{i,M+1}}{1-\pi_{i,M+1}}\right) = \hat{\theta}_0 + \sum_{j=1}^{M}\hat{\theta}_j\left(LP_{i,1}v_{1,j} + \ldots + LP_{i,M}v_{M,j}\right)$$
$$= \hat{\theta}_0 + \sum_{j=1}^{M}\sum_{r=1}^{M}\hat{\theta}_j v_{r,j} LP_{i,r}.$$

Model redevelopment

Let $\boldsymbol{X}_{M+1}$ denote the $n_{M+1} \times P$ matrix of predictors in the local population with associated binary outcomes $\boldsymbol{Y}_{M+1}$. Then the redevelopment approaches aim to derive a new CPM of the form

$$\log\left(\frac{\pi_{i,M+1}}{1-\pi_{i,M+1}}\right) = \hat{\beta}_{0,M+1} + \sum_{p=1}^{P}\hat{\beta}_{p,M+1}x_{i,p}$$

for $i = 1, \ldots, n_{M+1}$, model intercept, $\hat{\beta}_{0,M+1}$, and coefficients, $\hat{\beta}_{p,M+1}$, thereby disregarding the existing CPMs. In this study, two strategies of redevelopment were considered; namely, backwards selection using Akaike Information Criterion (AIC) and penalised maximum likelihood estimation (ridge regression). The AIC of a model is defined as $2k - 2\log(\text{L})$, where k is the number of estimated parameters and L is the maximum likelihood value. Backwards selection under AIC proceeds by starting with the full model (i.e. all available predictors) and iteratively removing predictors until the model that minimises the AIC is obtained. Conversely, ridge regression estimates the coefficients from the full model by maximising the following penalised log-likelihood function

$$l^*\left(\hat{\beta}_{M+1}\right) = \left(\sum_{i=1}^{n_{M+1}}\left\{y_i \log\left(\pi_{i,M+1}\right) + (1-y_i)\log\left(1-\pi_{i,M+1}\right)\right\}\right)$$
$$-\lambda\left(\sum_{p=1}^{P}\left(\hat{\beta}_{p,M+1}\right)^2\right).$$

Thus, the penalty shrinks the model coefficients towards zero, with λ controlling the degree of penalisation; cross-validation was used to select λ that minimised the deviance function.

Simulation design: general overview

Figure 1 visualises the simulation design. The simulation procedure generated both Normally distributed continuous predictors and Bernoulli distributed binary predictors, each within clusters of serially correlated variables

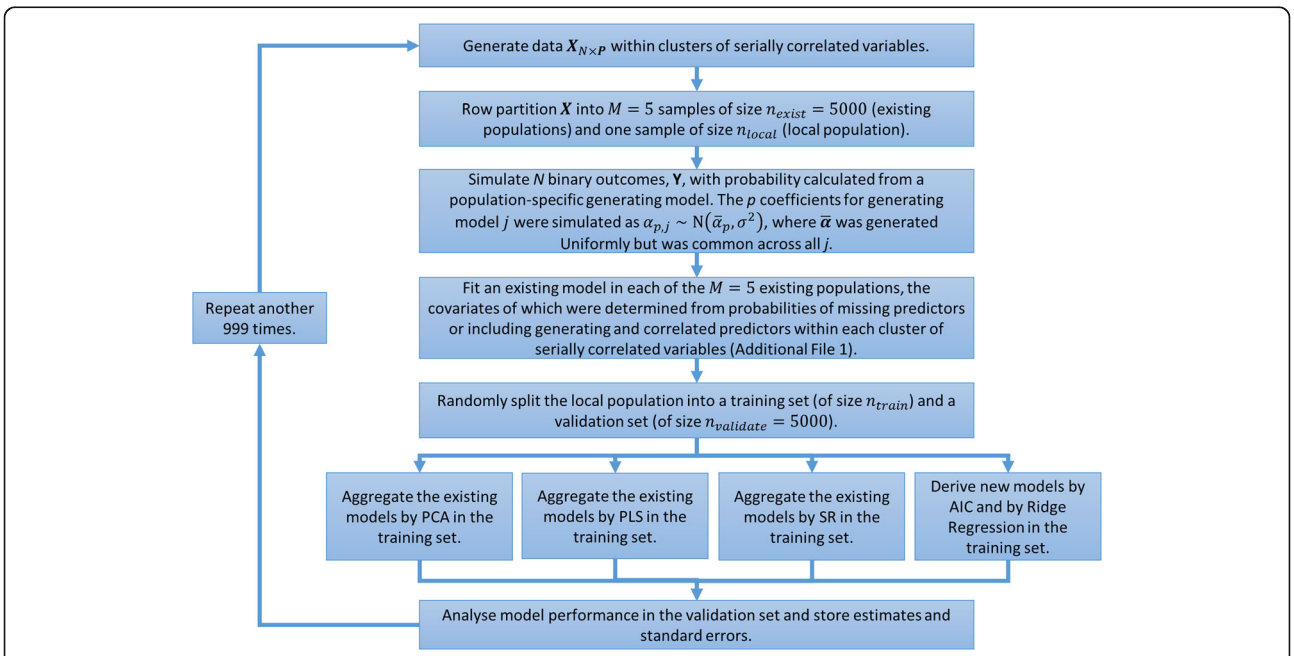

Fig. 1 Simulation Procedure: A pictorial representation of the simulation procedure for a given value of population heterogeneity, σ, and a given development sample size, n_{train}. This process was then repeated across all combinations of σ and n_{train}

to represent multiple risk factors that measure similar patient characteristics. Such data were row partitioned into $M = 5$ distinct subsets of size $n_{exist} = 5000$ representing five "existing populations", and one subset of size n_{local} representing the "local population". The $M = 5$ existing populations were each used to fit an existing logistic regression CPMs representing those available from the literature, with each CPM including a potentially overlapping subset of risk predictors (see Additional file 1: Table S1 for details of predictor selection for the existing CPMs). The single local population was randomly split into a training and validation set, of sizes n_{train} and $n_{validate}$, respectively (i.e. $n_{local} = n_{train} + n_{validate}$). The training set was used for model aggregation using SR, PCA and PLS in addition to redevelopment using AIC and ridge regression. Datasets frequently only collect a subset of the potential risk factors; to recognise this, exactly those predictors that were included in any of the five existing CPMs were considered candidates during redevelopment. Between simulations n_{train} was varied through (150, 250, 500, 1000, 5000, 10000); the validation set was reserved only to validate the models with $n_{validate}$ fixed at 5000 observations. Whilst it is unlikely that local populations would have access to such a large validation set, this was selected here to give sufficient event numbers for an accurate assessment of model performance [21–23]. Additionally, although bootstrapping methods are preferable to assess model performance in real-world datasets, the split-sample method was employed here for simplicity and clear illustration of the methods [24].

Binary responses were simulated in all populations with probability calculated from a population-specific generating logistic regression model, which included a subset of the simulated risk predictors. The coefficients of each population-specific generating model were sampled from a normal distribution, with a common mean across populations and variance σ. Here, higher values of σ induced greater differences in predictor effects across populations and thus represented increasing between-population-heterogeneity. For each of the aforementioned values of n_{train}, simulations were run with σ values of (0, 0.125, 0.25, 0.375, 0.5, 0.75, 1).

Across every combination of σ and n_{train}, the simulation was repeated over 1000 iterations as a compromise between estimator accuracy and computational time. The simulations were implemented using R version 3.2.5 [25]. The following packages were used in the simulation: "pROC" [26] to calculate the AUC of each model, "plsRglm" [27] to fit the PLS models and the "cv.glmnet" function within the "glmnet" package for deriving a new model by cross-validated ridge regression [28]. The authors wrote all other code, which is available in Additional file 1.

Simulation design: data-generating mechanisms

In practice, modellers could define any one risk factor through different but potentially related variables and multiple risk factors within a model could be correlated. Hence, the simulation procedure generated risk predictors within clusters of serially correlated variables. Specifically, $P = 50$ predictors were generated within 10 clusters, so that each cluster included $K = 5$ predictors. Predictors had serial correlation within each cluster but were independent between clusters. To represent common real data structures, the simulation generated clusters of binary and continuous predictors in an approximately 50/50 split, with the 'type' of each cluster decided at random before each simulation. For simplicity, clusters did not 'mix' binary and continuous variables. If $\boldsymbol{X}_{N\times P}$ denotes the $N \times$ P matrix of predictors (where N is the cumulative sample size across all populations) and ρ denotes the within-cluster correlation, then the process to generate the predictors was adapted from previous studies [29] as follows:

1. If cluster κ includes only continuous predictors then simulate N realisations of the predictors at the 'start' of the cluster as

   $$X_p \sim \text{Normal}(0, 1),$$

   and simulate the remaining *K-1* correlated predictors as

   $$X_p \sim \rho X_{p-1} + \sqrt{(1-\rho^2)}\Psi,$$

   where $\Psi \sim \text{Normal}(0, 1)$.
2. Else, if cluster κ includes only binary predictors, we generate them as latent Normal. Specifically, simulate N realisations of the predictors at the 'start' of each cluster as

   $$X_p \sim \text{Normal}(0, 1),$$

   and simulate the remaining *K-1* correlated predictors as

   $$X_p \sim \begin{cases} X_{p-1} \text{ with prob. } \rho \\ \Psi \text{ with prob. } 1-\rho \end{cases}$$

   where $\Psi \sim \text{Normal}(0, 1)$. Each variable in the cluster was then dichotomized to give a pre-defined cluster-specific event rate between 10 and 50%, which are values frequently reported in observational datasets.
3. Repeat steps 1 to 2 across all $\kappa = 10$ clusters.

Sensitivity analyses across a range of within-cluster correlations, $\rho \in [0, 0.99]$ showed that the results were qualitatively similar; the results given are for $\rho = 0.75$.

Binary responses for individuals $i = 1, \ldots, n_j$ in population j were sampled from a population-specific generating logistic regression model with $P(Y_{i,j} = 1) = q_{i,j}$, where

$$\log\left(\frac{q_{i,j}}{1-q_{i,j}}\right) = \alpha_{0,j} + \sum_{p=1}^{P=50} \alpha_{p,j} x_{i,p}$$

with intercept $\alpha_{0,j}$ and generating coefficients $\alpha_{1,j}, \ldots, \alpha_{50,j}$. If $\overline{\boldsymbol{\alpha}}$ represents the vector of mean predictor effects across all populations, then the simulation mechanism in each population j and generating parameter $p = 1, \ldots, 50$ was

$$\alpha_{p,j} \sim \begin{cases} \mathrm{N}\left(\overline{\alpha}_p, \sigma^2\right) & \text{if } p \equiv 1 \ (\mathrm{mod}\ \mathrm{K} = 5) \\ 0 & \text{otherwise} \end{cases}$$

The $p \equiv 1 \pmod{K = 5}$ condition implies (without loss of generality) that in each population, all non-zero generating coefficients were those at the 'start' of each cluster. Further, such a simulation procedure induced between-population-heterogeneity by applying random variation to the mean predictor-effects ($\overline{\boldsymbol{\alpha}}$), which was controlled through the value of σ that was introduced above. To represent coefficients frequently reported in published models, $\overline{\boldsymbol{\alpha}}$ was sampled in each simulation as follows:

$$\overline{\alpha}_p \sim \begin{cases} \mathrm{Uniform}(0.80,\ 1.6) & \text{if parameter } p \text{ is binary} \\ \mathrm{Uniform}(0.08,\ 0.1) & \text{if parameter } p \text{ is continuous} \end{cases}$$

In addition, baseline risk undoubtedly differs between populations and, as such, each intercept $\alpha_{0,j}$ was selected to give an average pre-defined event rate of 20% plus random variation. All simulations were repeated with an event rate of 50%, reflecting a 1-to-1 case-control study. A sensitivity analysis was undertaken where the magnitude of $\overline{\boldsymbol{\alpha}}$ was different across the generating predictors (see Additional file 1 for details), but the results were qualitatively similar as those presented here and so are omitted.

Simulation design: performance measures

For each iteration within a given simulation scenario, the mean squared error (MSE) between the predicted risk from each aggregate/new model and the actual risk from the generating model were calculated across all samples in the validation set. That is, for model m we have, $\mathrm{MSE_m} = \frac{1}{n_{validate}} \sum_{i=1}^{n_{validate}} \left(\hat{\pi}_{i,m} - q_i\right)^2$, where $\hat{\pi}_{i,m}$ is the predicted risk from model m for observation i in the validation set and q_i is the generating model risks for this observation. Similarly, the MSE was calculated between the estimated coefficients of each aggregate/new model and the generating coefficients (i.e. $\mathrm{MSE_m} = \frac{1}{P} \sum_{p=1}^{P} \left(\hat{\beta}_{p,m} - \alpha_{p,M+1}\right)^2$, where $\hat{\beta}_{p,m}$ is the estimated p^{th} coefficient from model m and $\alpha_{p,M+1}$ is the p^{th} generating coefficient in the local population). Additionally, the calibration and discrimination of each aggregate/new model were calculated in the validation set. The calibration was quantified with a calibration intercept and slope, where values of zero and one respectively represent a well-calibrated model [30]. Discrimination was evaluated by the area under the ROC curve (AUC). All performance measures were averaged across iterations and the empirical standard errors were calculated.

Results

Simulated between-population heterogeneity

To gain a practical understanding of the between-population-heterogeneity generated by increasing values of σ, for all simulated parameters the difference between the largest coefficient and smallest coefficient across populations was calculated and summarised (Table 1); such values were compared with corresponding values from the surgical models. Coefficient differences over the LES, ESII, STS and German AV represent heterogeneity across cardiac surgery risk models each developed across multiple countries. Coefficient differences over these models closely matched those generated with σ = 0.25 or σ = 0.375. Conversely, LES and ESII are two models that were developed on very similar cohorts of patients; here, the coefficient differences most closely match those generated by σ = 0.125. Similarly, the average standard deviation of the coefficients across the LES, ESII, STS and German AV models was 0.33 (closely matching σ = 0.375), while that between the LES and ESII was 0.26 (closely matching σ = 0.25). Together, this suggests that σ values between 0

Table 1 Summary measures of the difference in generating coefficients values across all simulated populations

	σ							LES, ESII, STS, German AV	LES, ESII
	0	0.125	0.25	0.375	0.5	0.75	1		
Lower Quartile[a]	0	0.29	0.58	0.86	1.16	1.74	2.31	0.37	0.14
Median[a]	0	0.31	0.63	0.95	1.27	1.90	2.52	0.57	0.31
Mean[a]	0	0.32	0.63	0.95	1.27	1.90	2.53	0.70	0.37
Upper Quartile[a]	0	0.34	0.68	1.03	1.38	2.06	2.74	0.85	0.54
SD[b]	0	0.12	0.24	0.36	0.48	0.71	0.95	0.33	0.26

The values for the LES, ESII, STS and German AV are approximate, since variable definitions vary between CPMs

[a]: values represent summary measures across all iterations of the average difference between the largest coefficient and smallest coefficient across populations

[b]: the average standard deviation (SD) of each coefficient across all populations. All values aim to guide the between-population heterogeneity induced through different σ values

and 0.375 likely represent the majority of clinical situations, with σ values greater than 0.5 arguably rare in practice.

Mean square error

For training set samples of 500 or less and when $\sigma \leq 0.25$, all three aggregation approaches resulted in predicted risks that had smaller mean square error and lowered the variance component of the error compared with redevelopment (Additional file 1: Table S2). Similarly, for training sample sizes less than or equal to 500, SR had estimated coefficients with consistently smaller mean square error with lower standard error than the redevelopment approaches, with the exception of the two highest values of σ (Additional file 1: Table S3). Conversely, for training samples of 1000 or more, developing a new model by ridge regression provided parameter estimates with at least equivalent MSE to the aggregation methods.

Model calibration

The calibration intercepts for all the aggregate/new models were not significantly different from zero in the validation set across all simulations (Fig. 2). Across all values of σ and for training set sizes smaller than 1000, the calibration slope of the AIC derived model was significantly below one indicating overfitting, while that for ridge regression was higher than one, indicating slight over-shrinkage on the parameters (Fig. 3). Conversely, the three aggregate models had a calibration slope not significantly different from one in any scenario, with the exception of PCA in the smallest sample sizes.

Model discrimination

When $\sigma \leq 0.125$ and for training sets of 250 or fewer, the AUC of SR was significantly higher than that of both redevelopment approaches (Fig. 4). Although the 95% confidence intervals overlapped, when $\sigma < 0.25$ and the training set was less than 1000 observations, the AUC of the two newly derived models (AIC/ridge) were less than that of the aggregate approaches (Additional file 1: Table S4). For instance, when $\sigma = 0$, the AUC of SR was higher than that of ridge regression in 988, 968, 821, 498, 56 and 19 out of the 1000 iterations for training set sizes 150, 250,

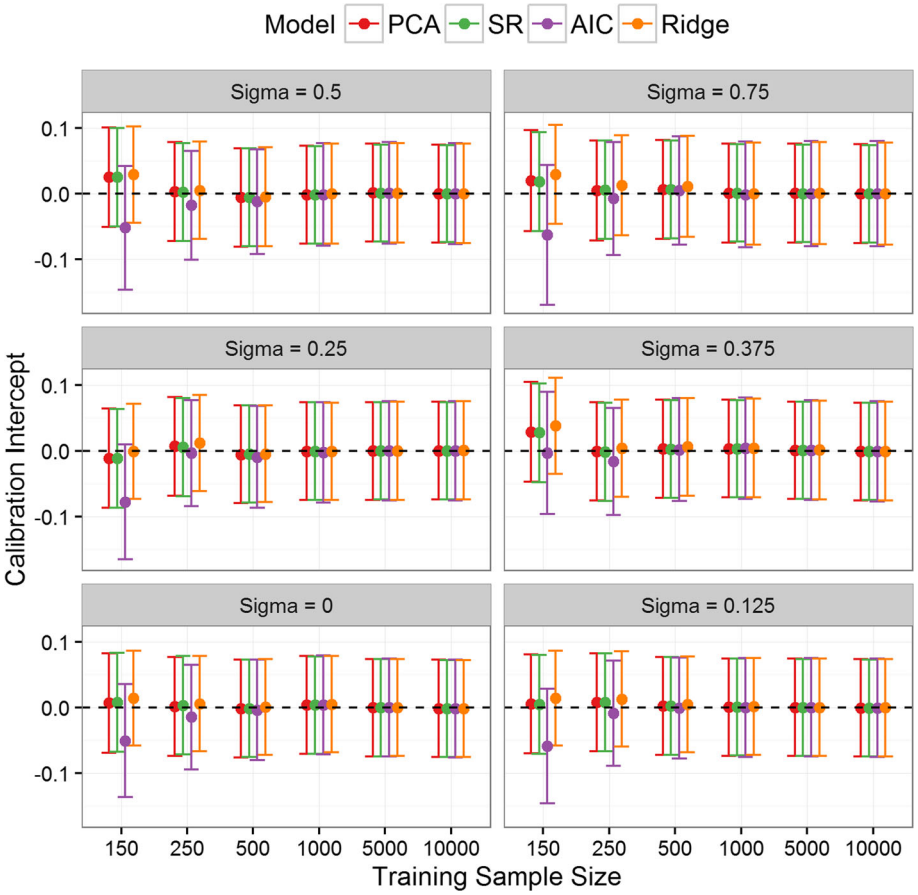

Fig. 2 Calibration Intercept: Calibration intercept in the validation set for SR, PCA and the two newly derived models across all simulation situations. The PLS results were nearly identical to SR/PCA and so are omitted for clarity. Note: $\sigma = 1$ was removed from the plot for clarity since the results quantitatively similar to $\sigma = 0.75$

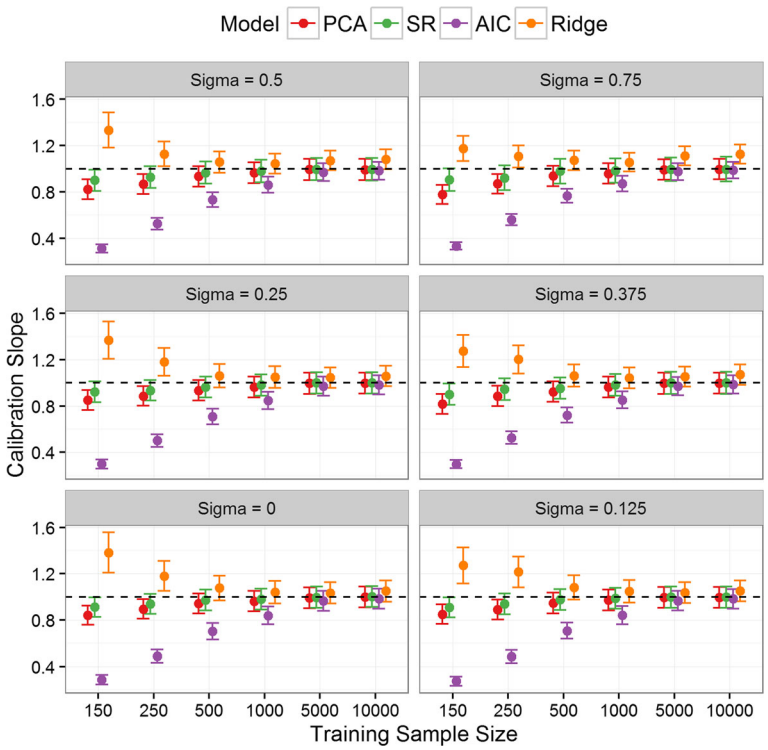

Fig. 3 Calibration Slope: Calibration slope in the validation set for SR, PCA and the two newly derived models across all simulation situations. The PLS results were nearly identical to SR/PCA and so are omitted for clarity. Note: $\sigma = 1$ was removed from the plot for clarity since the results quantitatively similar to $\sigma = 0.75$

500, 1000, 5000 and 10000, respectively. Hence, given training set samples of less than 500 and very similar populations, SR provides consistently higher AUC than redevelopment by either ridge or backwards selection.

Modelling strategy recommendations

A framework that compared modelling strategies of redevelopment and aggregation was developed. For redevelopment, ridge regression was always recommended over AIC since the former more appropriately accounted for low training set sizes. Likewise, all three aggregation approaches performed comparably and so SR was considered here due to the simplicity of implementation. Hence, on comparing ridge regression to SR across all simulation scenarios, if one of the models was well calibrated (calibration intercepts and slopes significantly close to zero and one, respectively) and had significantly higher AUC than the other model, then that modelling strategy was recommended. Conversely, if both models were well calibrated but the AUCs were not significantly different, then a recommendation of "Either" was given. Finally, if one of the models was miscalibrated then the other (calibrated) modelling strategy was recommended.

When the size of the training set was less than 500, then aggregating previously published models by SR was recommended (Table 2). Conversely, developing a new model by ridge regression was recommended in situations where $\sigma > 0.375$ and the size of the training set was at least 1000 observations. Between these scenarios, both aggregation methods and redevelopment methods provided indistinguishable performance. Similar recommendations were given when the average event prevalence was increased to 50% (Table 3).

Discussion

This study demonstrates that aggregating multiple published CPMs is a useful derivation strategy, particularly when there are limited data available. Stacked regression was a simple yet effective aggregation method, which resulted in predictions and parameter estimates with lowest MSE given low sample sizes and low between-population-heterogeneity. These results are consistent with previous studies [9]. Conversely, AIC derived models were miscalibrated when the training set sample size was between 150 and 500, confirming that small samples lead to overfitting in new regression estimates [8, 31, 32]. Ridge regression, which is a similar concept to parameter shrinkage, mitigated overfitting but was potentially susceptible to slight over-shrinkage. Redevelopment only resulted in a model with better performance than the aggregation methods when there was a large amount of training data or the existing CPMs were significantly heterogeneous.

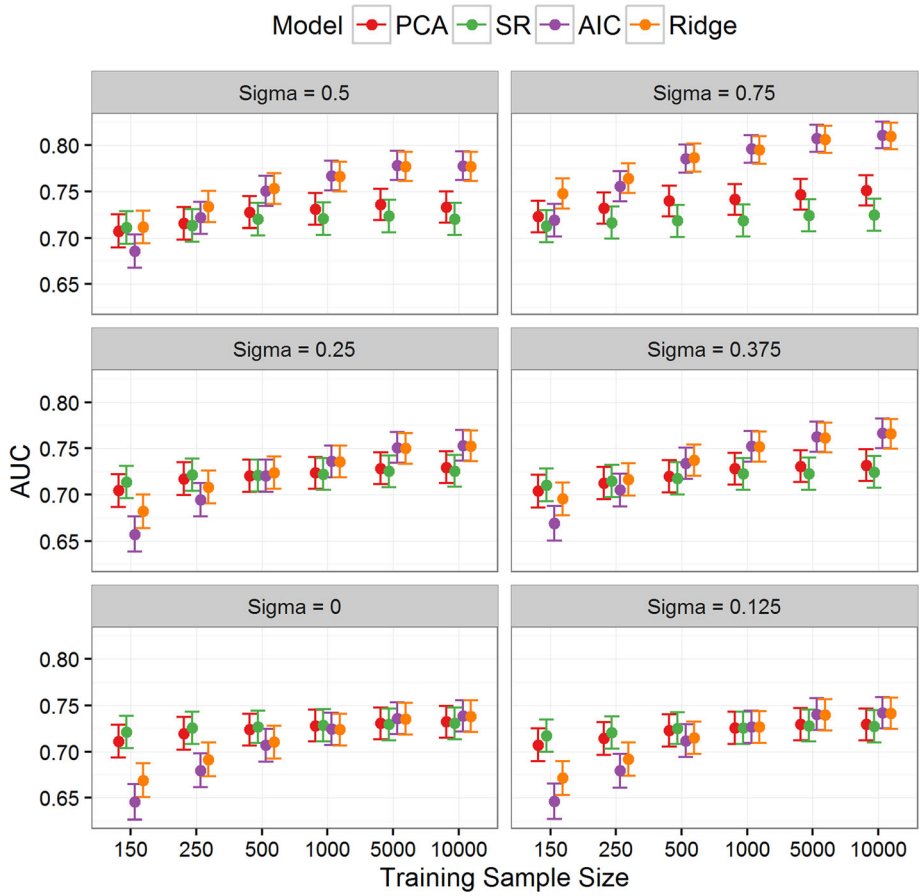

Fig. 4 Discrimination: Area under receiver operating characteristic curve (AUC) in the validation set for SR, PCA and the two newly derived models across all scenarios. The PLS results were nearly identical to SR/PCA and so are omitted for clarity. Note: σ = 1 was removed from the plot for clarity since the results quantitatively similar to σ = 0.75

Previous methodological research around incorporating existing CPMs has focussed on updating a single existing model to the new population of interest [7, 8, 10, 33]. These techniques range from model recalibration to the additional of new risk factors and have been shown to provide improved performance over deriving new prediction models, especially when only small datasets are available [8]. However, updating techniques only adapt one previous model to the current data. In this sense, the concept of model aggregation is analogous to meta-analysis since it aims to synthesise all previous research in the development of the CPM. Moreover, CPMs often perform poorly in high-risk patients. If there are relatively low proportions of high-risk patients in a given population, then the

Table 2 Modelling strategy recommendations when the mean incidence of adverse outcome was 20%

σ	n_{train}					
	150	250	500	1000	5000	10000
0	SR	SR	Either	Either	Either	Either
0.125	SR	SR	Either	Either	Either	Either
0.25	SR	SR	Either	Either	Either	Either
0.375	SR	SR	Either	Either	Ridge	Ridge
0.5	SR	SR	Either	Ridge	Ridge	Ridge
0.75	SR	SR	Ridge	Ridge	Ridge	Ridge
1.00	SR	SR	Ridge	Ridge	Ridge	Ridge

Table 3 Modelling strategy recommendations when the mean incidence of adverse outcome was 50%

σ	n_{train}					
	150	250	500	1000	5000	10000
0	SR	SR	Either	Either	Either	Either
0.125	SR	SR	Either	Either	Either	Either
0.25	SR	SR	Either	Either	Either	Either
0.375	SR	SR	Either	Ridge	Ridge	Ridge
0.5	SR	Ridge	Ridge	Ridge	Ridge	Ridge
0.75	Ridge	Ridge	Ridge	Ridge	Ridge	Ridge
1.00	Ridge	Ridge	Ridge	Ridge	Ridge	Ridge

development/ update of a CPM to this population can result in such high-risk, poorly predicted patients becoming more prevalent since parameter estimates occur for the population average. In such situations, one should pay close attention to the residuals of the model; machine-learning methods such as Boosting are a formal approach to this.

Since the aim of this study was to examine the benefits of aggregation over independently deriving a CPM, this study compared each approach separately to solely extract the benefit of either method. However, meta-analysis methods that simultaneously aggregate and redevelop CPMs have been proposed [18, 34, 35]; utilising existing CPMs, expert knowledge and new data optimally requires further research. For instance, risk factors may not be common across existing CPMs, which could lead to bias if one is interested in simultaneously aggregating and redeveloping CPMs in the local population [36]. Previous methodology of CPM meta-analysis with individual patient data has largely been limited to assuming that models share similar risk predictors [10, 18]. Conversely, SR, PCA and PLS relax this assumption [9]. Indeed, the simulation design of this study allowed the existing CPMs to be heterogeneous in their risk predictor set.

Nevertheless, there are potential problems of aggregating CPMs that require attention. Firstly, each existing CPM aims to predict the same outcome and most include very similar subset of predictors, thus inducing a high level of correlation between the multiple CPMs. Although the weights in SR are restricted to be non-negative to avoid situations of negative coefficients caused by inclusion of two correlated models, further work examining the collinearity issues is required [10]. Secondly, differences in risk factor definitions between existing CPMs could potentially weaken the performance of SR, PCA or PLS. The current study aimed to replicate this practical limitation by generating predictors within clusters of correlated variables; here, given a moderate degree of correlation between the multiple similarly defined risk factors, the aggregation methods still performed well. Finally, datasets across populations frequently collect different variables, potentially meaning a variable included in an existing CPM is not available in the local population. In such circumstances of systematically missing covariates, it is unclear how one should calculate the linear predictor for patients in the new local population [37]. If systematically missing risk factors are not handled appropriately, then the aggregate CPM could be biased.

The main strength of this work is that we perform a simulation study under a range of realistic scenarios and consider multiple performance measures, thereby allowing a comprehensive and systematic examination of the aggregation approaches. Conversely, the main limitation is that we simulate only a crude reflection of real between-population-heterogeneity. Over-arching variance of model parameters does not necessarily reflect the complex differences in data-generating processes that may vary between populations. However, without a comprehensive set of joint probability distributions for the covariates of a given model, accurately modelling population heterogeneity is difficult to achieve. Hence, confirmation of our findings will be required from studies in observational datasets. A further limitation is that publication bias is known to impact prognostic research [38], but its effects were not analysed in this study; such bias would lead to overestimation of aggregated regression coefficients. Finally, the aggregation methods assume that each population is a random sample from an over-arching common population. The data-generating mechanisms in this simulation directly matched this assumption by simulating generating model coefficients as a random sample from a common distribution. Similarly, the distributions of the risk predictors were assumed the same between populations.

Overall, the current work suggests a framework of modelling strategy when developing a model for a local/ defined population. In practice, an estimate of the between-population-heterogeneity could be approximated by examining the differences in coefficients between existing CPMs, exploiting clinical knowledge between networks of modelling teams and by examining the distribution of the linear predictors between populations [39]. In many practical scenarios, the variability between populations will be low; thus the situations of $\sigma = 0$ to $\sigma = 0.375$ in the current study likely closely represent clinical practice. If the size of the local data is <10% of that the existing CPMs were derived on, and if the multiple populations share clinically similar demographic and procedural characteristics, then we recommend aggregating existing models. Secondly, if the size of local data matches or exceeds that of existing model derivations, then deriving a new CPM could be appropriate, although the existing CPMs could still provide useful prior information about likely covariate-outcome associations. Finally, in all other circumstances, one should consider either aggregation, redevelopment or a combination of the two [18]. Here, the sample sizes relative to the number of predictors per event [31], the estimated population heterogeneity, the quality of the existing CPMs and the availability of variables should drive the chosen method.

Conclusions

Aggregating existing CPMs is beneficial in the development of a CPM for healthcare predictions in defined populations. In the majority of situations, modellers should consider existing CPMs before developing models anew, with their aggregation potentially providing optimal

performance given low sample sizes relative to that of previous model derivations. Deriving a new CPM independent of previous research was only recommended in the unusual situation of having more data available than used to derive existing models, or a local context that is markedly different to those of existing CPMs.

Additional file

Additional file 1: Details of the Partial Least Squared regression methodology, additional details of the simulation design, additional tables and the simulation R code. (DOCX 49 kb)

Abbreviations

CPM: Clinical prediction model; ESII: EuroSCORE II; German AV: German Aortic Valve Model; LES: Logistic EuroSCORE; MSE: Mean Square Error; PCA: Principal component analysis; PLS: Partial least squares; SR: Stacked regression

Acknowledgements

Not applicable.

Funding

This work was funded by the Medical Research Council through the Health e-Research Centre, University of Manchester [MR/K006665/1].

Availability of data and material

The data on which the conclusions of this manuscript rely can be reproduced using the R code available in the online Additional file 1.

Authors' contributions

GPM and MS designed the study and drafted the initial version of the manuscript. GPM wrote the simulation R code. GPM, MAM, NP, IB and MS analysed/interpreted the results and revised the manuscript critically for important intellectual content. All of the authors approved the final version and agreed to be accountable for all aspects of the work.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Health e-Research Centre, University of Manchester, Vaughan House, Portsmouth Street, M13 9GB Manchester, UK. [2]Keele Cardiovascular Research Group, Keele University, Stoke-on-Trent, UK. [3]NIHR Greater Manchester Primary Care Patient Safety Translational Research Centre, University of Manchester, Manchester, UK.

Received: 17 August 2016 Accepted: 15 December 2016
Published online: 06 January 2017

References

1. 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 Mak. 2012;32:E1–E10.
2. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, Lassale CM, Siontis GCM, Chiocchia V, Roberts C, Schlüssel MM, Gerry S, Black JA, Heus P, van der Schouw YT, Peelen LM, Moons KGM. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416.
3. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605.
4. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. BMJ. 2009;338:b604.
5. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606.
6. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140.
7. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol. 2008;61:76–86.
8. 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;23:2567–86.
9. Debray TPA, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW, Moons KGM. Meta-analysis and aggregation of multiple published prediction models. Stat Med. 2014;33:2341–62.
10. Su T-L, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods for clinical prediction models. Stat Methods Med Res. 2016. doi:10.1177/0962280215626466.
11. Altman DG. Prognostic Models: A Methodological Framework and Review of Models for Breast Cancer. Cancer Invest. 2009;27:235–43.
12. Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, Lockowandt U. EuroSCORE II. Eur J Cardio-Thoracic Surg. 2012;41:734–45.
13. Roques F. The logistic EuroSCORE. Eur Heart J. 2003;24:882.
14. O'Brien SM, Shahian DM, Filardo G, Ferraris VA, Haan CK, Rich JB, Normand S-LT, DeLong ER, Shewan CM, Dokholyan RS, Peterson ED, Edwards FH, Anderson RP. The Society of Thoracic Surgeons 2008 Cardiac Surgery Risk Models: Part 2—Isolated Valve Surgery. Ann Thorac Surg. 2009;88:S23–42.
15. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, Normand S-LT, DeLong ER, Shewan CM, Dokholyan RS, Peterson ED, Edwards FH, Anderson RP. The Society of Thoracic Surgeons 2008 Cardiac Surgery Risk Models: Part 3—Valve Plus Coronary Artery Bypass Grafting Surgery. Ann Thorac Surg. 2009;88:S43–62.
16. Kotting J, Schiller W, Beckmann A, Schafer E, Dobler K, Hamm C, Veit C, Welz A. German Aortic Valve Score: a new scoring system for prediction of mortality related to aortic valve procedures in adults. Eur J Cardio-Thoracic Surg. 2013;43:971–7.
17. Breiman L. Stacked Regression. Mach Learn. 1996;24:49–64.
18. 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;31:2697–712.
19. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24:417–41.
20. Merz CJ, Pazzani MJ. A Principal Components Approach to Combining Regression Estimates. Mach Learn. 1999;36:9–32.
21. Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58:475–83.
22. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35:214–26.
23. Peek N, Arts DGT, Bosman RJ, van der Voort PHJ, de Keizer NF. External validation of prognostic models for critically ill patients required substantial sample sizes. J Clin Epidemiol. 2007;60:491–501.
24. 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. doi:10.1177/0962280214558972.
25. R Core Team R: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing 2016. [R Foundation for Statistical Computing]
26. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
27. Bertrand F, Meyer N, Maumy-Bertrand M. Partial Least Squares Regression for Generalized Linear Models. 2014.
28. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33:1–22.
29. Sperrin M, Jaki T. Recovering Independent Associations in Genetics: A Comparison. J Comput Biol. 2012;19:978–87.
30. Cox D. Two further applications of a model for binary regression. Biometrika. 1958;45:562–5.

31. Steyerberg E. Stepwise Selection in Small Data Sets A Simulation Study of Bias in Logistic Regression Analysis. J Clin Epidemiol. 1999;52:935–42.
32. Steyerberg EW, Eijkemans MJC, Harrell FE, Habbema JDF. Prognostic Modeling with Logistic Regression Analysis: In Search of a Sensible Strategy in Small Data Sets. Med Decis Mak. 2001;21:45–56.
33. Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: A review. J Clin Epidemiol. 2008;61:1085–94.
34. Steyerberg EW, Eijkemans MJC, Van Houwelingen JC, Lee KL, Habbema JDF. Prognostic models based on literature and individual patient data in logistic regression analysis. Stat Med. 2000;19:141–60.
35. Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. J Clin Epidemiol. 2007;60:431–9.
36. Yoneoka D, Henmi M, Sawada N, Inoue M. Synthesis of clinical prediction models under different sets of covariates with one individual patient data. BMC Med Res Methodol. 2015;15:101.
37. Held U, Kessels A, Garcia Aymerich J, Basagaña X, ter Riet G, Moons KGM, Puhan MA. Methods for Handling Missing Variables in Risk Prediction Models. Am J Epidemiol. 2016. doi:10.1093/aje/kwv346.
38. Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. BMJ. 2009;339:b4184.
39. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68:279–89.