Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model

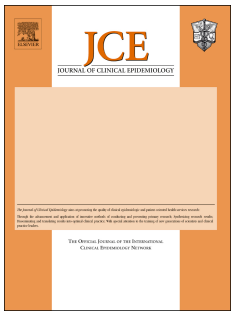
Snell, Kym; Hua, Hairui; Debray, T.p.a.; Ensor, J.; Look, M.p.; Moons, K.g.m.; Riley, R.d.

DOI:
10.1016/j.jclinepi.2015.05.009

Document Version
Peer reviewed version

Citation for published version (Harvard):
Snell, K, Hua, H, Debray, TPA, Ensor, J, Look, MP, Moons, KGM & Riley, RD 2015, 'Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model', *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2015.05.009

Accepted Manuscript



Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model

K.I.E. Snell, H. Hua, T.P.A. Debray, J. Ensor, M.P. Look, K.G.M. Moons, R.D. Riley

PII: S0895-4356(15)00230-9

DOI: 10.1016/j.jclinepi.2015.05.009

Reference: JCE 8882

To appear in: *Journal of Clinical Epidemiology*

Received Date: 14 October 2014

Revised Date: 5 May 2015

Accepted Date: 8 May 2015

Please cite this article as: Snell K, Hua H, Debray T, Ensor J, Look M, Moons K, Riley R, Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model, *Journal of Clinical Epidemiology* (2015), doi: 10.1016/j.jclinepi.2015.05.009.

Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model

Snell KIE[1], Hua H[2], Debray TPA[3,4], Ensor J[5], Look MP[6], Moons KGM[3,4], Riley RD[5]*

Contact details:

* corresponding author:

Professor of Biostatistics; e- mail: r.riley@keele.ac.uk;

Tel: +44 (0) 1782 733905 Fax: +44 (0) 1782 734719

[1] School of Health and Population Sciences, University of Birmingham, Edgbaston, Birmingham, UK. B15 2TT

[2] School of Mathematics, University of Birmingham, Edgbaston, Birmingham, UK. B15 2TT

[3] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

[4] Dutch Cochrane Centre, University Medical Center Utrecht, Utrecht, The Netherlands

[5] Research Institute of Primary Care and Health Sciences, Keele University, Staffordshire, ST5 5BG

[6] Dept. Medical Oncology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, the Netherlands.

Acknowledgements:

Richard D Riley and Joie Ensor are supported by funding from a multivariate meta-analysis grant from the MRC Methodology Research Programme (grant reference number: MR/J013595/1). Kym I.E. Snell was supported by funding from the MRC Midlands Hub for Trials Methodology Research, at the University of Birmingham (Medical Research Council Grant ID G0800808). Karel G.M. Moons acknowledges financial contribution by the Netherlands Organisation for Scientific Research (project 9120.8004 and 918.10.615). We thank the investigators of the uPA/PAI-1 prognostic breast cancer pooled study to use their anonymized patient data sets. We thank two anonymous reviewers for constructive comments that helped improve the article considerably.

We gratefully acknowledge the investigators for sharing of the anonymized individual participant data from the deep vein thrombosis (DVT) studies: A.J. Ten-Cate-Hoek, C. Kearon, H.R. Büller, H.C.van Weert, M.H. Prins, H.E. Stoffers, R. Oudega, K.G.M. Moons, R.E.G. Schutgens, R.A. Kraaijenhagen, D.B. Toll, D.R. Anderson, P.S. Wells, J.L. Elf, S.M. Bates, S.M. Stevens, S.C. Woller.

Abstract

Objective: Our aim was to improve meta-analysis methods for summarising a prediction model's performance when individual participant data are available from multiple studies for external validation.

Study design & setting: We suggest multivariate meta-analysis for jointly synthesising calibration and discrimination performance, whilst accounting for their correlation. The approach estimates a prediction model's average performance, the heterogeneity in performance across populations, and the probability of 'good' performance in new populations. This allows different implementation strategies (e.g. recalibration) to be compared. Application is made to a diagnostic model for deep vein thrombosis (DVT) and a prognostic model for breast cancer mortality.

Results: In both examples multivariate meta-analysis reveals that calibration performance is excellent on *average*, but highly heterogeneous across populations unless the model's intercept (baseline hazard) is recalibrated. For the cancer model, the probability of 'good' performance (defined by C-statistic≥0.7 and calibration slope between 0.9 and 1.1) in a new population was 0.67 with recalibration, but 0.22 without recalibration. For the DVT model, even with recalibration there was only a 0.02 probability of 'good' performance.

Conclusion: Multivariate meta-analysis can be used to externally validate a prediction model's calibration and discrimination performance across multiple populations, and to evaluate different implementation strategies.

Keywords

Risk prediction; prognostic model; individual participant data (IPD); multivariate meta-analysis; external validation; calibration; discrimination; heterogeneity; model comparison

What is new?

KEY FINDINGS

Given individual participant data (IPD) from multiple external validation studies, meta-analysis enables researchers to summarise prediction model performance, in terms of both average performance and consistency in performance across populations. It thereby allows different implementation strategies (e.g. recalibration) to be formally compared.

A *multivariate* meta-analysis approach should be used to jointly evaluate discrimination and calibration performance, whilst accounting for their correlation. This can be used within internal-external cross-validation (to also incorporate a model development phase), or when IPD from multiple studies are available for external validation of existing models.

WHAT THIS ADDS TO WHAT IS KNOWN:

Before implementation, risk prediction models require validation in data external to that used for model development. This is best achieved using IPD from multiple studies, so that model performance can be examined and quantified across multiple populations of interest. A good prediction model will have satisfactory performance on average across all external validation datasets, and crucially little or no between-study heterogeneity in performance.

Our examples show that a prediction model may have excellent *average* performance, but with heterogeneity (inconsistency) in performance across populations. Recalibration of the model's intercept term (or baseline hazard) in the intended population might reduce heterogeneity, and thereby improve the probability of acceptable model performance when applied in new populations.

WHAT IS THE IMPLICATION, WHAT SHOULD CHANGE NOW:

When IPD are available from multiple studies for external validation of a prediction model, researchers should use multivariate meta-analysis to jointly summarise calibration and discrimination performance, and to identify how best to implement the model in new populations.

1. Introduction

A crucial part of medical research is to develop risk prediction models. These aim to accurately predict disease and outcome risk in individuals [1-3], thereby informing clinical diagnosis and prognosis. For example, healthy individuals with a high predicted risk of future disease (e.g. cardiovascular events) may be advised to modify their lifestyle and behaviour choices (e.g. smoking, exercise), and diseased individuals may be grouped (e.g. stage of cancer) according to future outcome risk so that clinical decisions (such as treatment options, monitoring strategies) can be tailored accordingly. Two well-known examples are QRISK [4] and the Nottingham Prognostic Index [5]. They are typically implemented within a multivariable regression framework, such as logistic or Cox regression, which provides an equation to estimate an individual's risk based on values of multiple predictors (prognostic factors [6]) such as age, biomarkers, and genetic information

A key stage of prediction model research is *model development* [2]. This identifies important predictors and develops the risk prediction equation using an available dataset; it usually also examines the model's *apparent* performance in this same data, or uses *internal validation* techniques (such as bootstrap resampling) to examine and adjust for optimism in performance [7]. The next stage is *external validation* [8-10]. This uses data external to the model development data and its source, and examines whether the model predictions are accurate in another (but related) situation. The aim is to ascertain the model's generalisability to the intended populations for use [11], and to identify the best implementation strategy (e.g. recalibration of the intercept).

Unfortunately, most prediction research focuses on model development and there are relatively few external validation studies [12]. However, nowadays there is increasing access to multiple datasets, as evident in meta-analyses using individual participant data (IPD) from multiple studies [13, 14]. This provides an exciting opportunity to perform external validation on multiple occasions [15, 16]. Model development and external validation can even occur simultaneously, using an approach called internal-external cross-validation [17, 18]. This develops a model in all but one of the IPD studies, and then its external validity is immediately checked in the omitted study. This process is repeated across all rotations of the omitted study, to measure external validity in each distinct IPD study.

Given multiple external validation studies, meta-analysis methods are needed to synthesise and summarise model performance appropriately across the available populations. Van Klaveren et al.[16], Pennells et al.[15] and, within internal-external cross-validation, Royston et al.[17] consider approaches to summarise validation performance across multiple studies or clusters. These focus mainly on producing pooled estimates of *discrimination* performance; that is, a model's ability to distinguish correctly between patients with and without the outcome of interest. Researchers should also be interested in summarising *calibration* performance, which is the agreement between a model's predicted risk and the observed risk. Calibration is often ignored in external validation research [19], even though it is fundamental that observed and predicted risks should closely agree. Moreover, baseline risk may vary across study populations and so a model's implementation may need to be tailored to each population (often referred to as *recalibration*) in order to improve calibration performance in new populations.

In this article, we propose multivariate meta-analysis for jointly synthesising discrimination *and* calibration performance, whilst accounting for their correlation. This can be used within internal-external cross-validation (to also incorporate a model development phase), or when IPD from multiple studies are available for external validation of existing models. We show that the multivariate approach summarises a prediction model's average discrimination and calibration performance, and quantifies the heterogeneity in performance across populations. It also allows researchers to predict the potential calibration and discrimination of a model when it is applied to a new population, and can be used to estimate the probability of 'good' performance (as pre-defined by the user). Using two real examples, we illustrate how this enables researchers to compare the performance of different implementation strategies (e.g. recalibration of the intercept term) to help identify the best strategy for applying the model in practice.

The article now proceeds by introducing the proposed multivariate meta-analysis methodology for summarising and comparing validation performance (Section 2). Two clinical examples are then used to illustrate the approach (Section 3), one for diagnosis and one for prognosis, and we conclude with some discussion (Section 4).

2. Meta-analysis of predictive performance statistics from multiple external validation studies

External validation of a prediction model requires evaluation of its predictive performance, in terms of both calibration and discrimination. There are many statistical measures available for this purpose [1, 20]. Here we focus on those most commonly used: the C-statistic [20, 21] the D-statistic [22, 23], the calibration slope [1, 20], calibration-in-the-large [1], and the Expected/Observed number of events. These are defined in the Appendix. We focus here on how to meta-analyse such performance statistics when they are estimated in multiple external validation studies.

2.1 Obtaining suitable data for meta-analysis

The meta-analysis approach requires an estimate of each performance statistic of interest (e.g. C-statistic, calibration slope) from each external validation study. Given IPD, these can be calculated in each validation study using appropriate statistical methods, as described elsewhere [1, 20, 23]. However, meta-analysis also requires the variance-covariance matrix of the performance statistics in each study: in other words, the variance of each performance estimate and (for multivariate meta-analysis) the correlation between all pairs of estimates. A general approach to obtaining these is via non-parametric bootstrapping, as described in the Appendix.

2.2 Univariate random-effects meta-analysis

For clarity, before proposing our multivariate approach we firstly describe a univariate random-effects meta-analysis that is applicable separately to each performance measure of interest[24, 25]. In external validation study i let Y_{ij} be the estimate of the j^{th} performance statistic of interest, and let S_{ij}^2 be its sample variance (derived from bootstrapping and assumed known), then the univariate meta-analysis can be written as:

$$\begin{aligned} Y_{ij} &\sim \mathrm{N}\left(\mu_{ij}\,, S_{ij}^2\right) \\ \mu_{ij} &\sim \mathrm{N}\left(\mu_j\,, \tau_j^2\right) \end{aligned} \qquad \textbf{(1)}$$

Equation (1) assumes the Y_{ij} are normally distributed about the i^{th} study's true validation performance, μ_{ij}, and that the μ_{ij} are also normally distributed with an average of μ_j and a

between-study standard deviation of τ_j. There are several frequentist methods that can be used for estimation of a random-effects meta-analysis; here we use restricted maximum likelihood (REML)[26]. With the addition of prior distributions for unknown parameters, a Bayesian approach is also possible, for example using Gibbs sampling. An approximate 100(1-α)% confidence interval for the average performance, μ_j, is obtained by $\hat{\mu}_j$ ±1.96 SE($\hat{\mu}_j$), where SE($\hat{\mu}_j$) is the standard error of $\hat{\mu}_j$. White [27] proposed that SE($\hat{\mu}_j$) is inflated to account for the uncertainty in the estimated τ_j, and we implement this here.

Summarising consistency in model performance

On its own, $\hat{\mu}_j$ is an incomplete summary because it does not adequately summarise the consistency in performance across studies. Estimates such as I^2 (the percentage of the total variation in study estimates that is due to between-study heterogeneity [28]) and $\hat{\tau}_j^2$ are thus also helpful [29]. However, when evaluating performance statistics of a risk prediction model we are examining its generalisability, in other words its robustness when applied in new populations that differ from those it was developed in [11]. Thus, consistency is best expressed by a 100(1-α)% prediction interval for the performance of the model in a new population [24, 25]. This is derived by

$$\hat{\mu}_j \pm t_{\alpha,\, N-2}\sqrt{\hat{\tau}_j^2 \quad +\mathrm{V}(\hat{\mu}_j)} \qquad \textbf{(2)}$$

where $t_{\alpha,\, N-2}$ is the $100\left(1-\frac{\alpha}{2}\right)$% percentile of the t-distribution for N - 2 degrees of freedom (N = no. of studies), V($\hat{\mu}_j$) =SE($\hat{\mu}_j$)^2, and α is typically taken to be 0.05 to give a 95% interval. The use of a t-distribution, rather than a normal distribution, is used to account for the uncertainty in $\hat{\tau}_j^2$ [24]. The prediction interval thus indicates the performance expected in a new (external validation) study, similar to those included in the meta-analysis.

2.3 Multivariate meta-analysis

Our multivariate approach is an extension of equation (1)[30], and allows the *joint* synthesis of all predictive performance measures of interest from the i = 1 to N external validation studies, whilst accounting for their within- and between-study correlation. Let there be j = 1 to J measures of interest, and let $\mathbf{Y_i}$ be a vector containing the available J estimates ($Y_{i1}, Y_{i2}, \ldots, Y_{iJ}$) of the measures in the i^{th} validation study. The general multivariate meta-analysis model is as follows:

$$\begin{aligned} \mathbf{Y_i} \,|\, \boldsymbol{\theta}_\mathbf{i} &\sim \text{MVN}(\boldsymbol{\theta}_\mathbf{i}, \mathbf{S_i}) \\ \boldsymbol{\theta}_\mathbf{i} &\sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned} \qquad \textbf{(3)}$$

Here MVN denotes a multivariate normal distribution, $\boldsymbol{\theta}_\mathbf{i}$ contains the true underlying effects for the *J* performance measures for the i^{th} study, $\mathbf{S_i}$ is the within-study variance-covariance matrix for the i^{th} study (assumed known) containing the *J* variances of the estimates (in the diagonal: $S_{i1}^2, S_{i2}^2, \ldots, S_{iJ}^2$) and their covariances (in the off-diagonal; for example $\rho_{Wi(1,2)} S_{i1} S_{i2}$ is the within-study covariance for measures 1 and 2, where $\rho_{Wi(1,2)}$ is their within-study correlation caused by estimates derived from the same patients), $\boldsymbol{\mu}$ contains the *J* means for the measures of interest, and $\boldsymbol{\Sigma}$ is the between-study variance-covariance matrix containing the *J* between-study variances (in the diagonal: $\tau_1^2, \tau_2^2, \ldots, \tau_J^2$) and their between-study covariances (in the off-diagonal; e.g. the between-study covariance for measures 1 and 2 is $\rho_{B(1,2)} \tau_1 \tau_2$, where $\rho_{B(1,2)}$ is their between-study correlation induced by differences in study populations and settings). The number of rows in each vector is equal to the number of measures. In its simplest form with two measures of interest (e.g. C-statistic and calibration slope), equation (3) can be expressed as a bivariate meta-analysis (Appendix).

REML can again be used for estimation, although other options are available [30, 31]. Multivariate extensions to I^2 can also be calculated [26, 31], giving the fraction of the total variability due to between-study variability for each performance statistic (I_J^2).

Making joint inferences across multiple performance measures

After equation (3) is estimated, marginal confidence and prediction intervals for each performance measure can be obtained using the formulae given in the univariate section. However, by accounting for their correlation, the multivariate approach also enables joint inferences. For instance, extending equation (2) to a bivariate t-distribution with $k - 2$ degrees of freedom, one can obtain a *joint* 95% prediction region for two performance measures of interest (e.g. the C-statistic and the calibration slope) in a new population. Joint probabilistic inferences can also be made if we assume the multivariate t-distribution is an approximate posterior distribution (that is, we assume it is obtained from a Bayesian analysis with uninformative priors, and give it means, variances and covariances obtained from REML

estimation of equation (3) - see Supplementary material 1 for full details [32]). For example, one can derive the joint probability that the C-statistic will be above 0.7 *and* the calibration slope will be between 0.9 and 1.1 in a new population. A fully Bayesian approach can also be used to derive such posterior inferences by formally specifying prior distributions and combining them with the likelihood, then using, for example, Gibbs sampling to take samples from the exact posterior distributions. Riley et al.[33] describe the Bayesian approach to multivariate meta-analysis with IPD.

2.4 Comparing the predictive performance of different implementation strategies

When applying a prediction model to a new population, different implementation strategies might be used regarding the choice of model intercept (baseline hazard) (this is illustrated in Sections 3.1 and 3.2, and for example includes recalibration). Meta-analysis of performance statistics allows such implementation strategies to be formally compared. The aim is to identify an implementation strategy that, for each performance measure, has excellent performance on average (indicated by $\hat{\mu}_j$); small values of between-study heterogeneity (indicated by $\hat{\tau}_j$ and/or I_j^2); and a narrow prediction interval that suggests consistently good performance in new populations. Multivariate meta-analysis even allows the competing strategies to be ranked according to their overall performance: for example, according to the joint probability that, in a new population, the C-statistic will be above 0.7 *and* the calibration slope will be between 0.9 and 1.1. The strategy with the largest probability will be ranked first.

2.5 Meta-regression and examining covariates

Meta-analysis equation (3) can be extended to a multivariate meta-regression that includes study-level covariates to explain between-study heterogeneity, such as treatment policies, population characteristics (e.g. mean age), year of investigation, and length of follow-up. Competing implementation strategies can then be evaluated and compared for specific subgroups of studies (e.g. those done within the last few years, those with consistent treatment policies, those with the same case-mix, etc). This may help identify populations where model performance is satisfactory and others where it is inadequate, to inform the model's generalisability and applicability [11]. A nice example of a meta-regression to examine the impact of case-mix variation on model performance is given by Pennells et al. [15], who identify that studies with a higher standard deviation of age are strongly associated

with a higher C-statistic and D-statistic. Model performance can also be examined for patient-level covariates; for example, discrimination and calibration could be estimated for males and females separately. Equation (3) can then be applied to summarise each subgroup, or even the difference between subgroups.

3. Applied examples

We now illustrate the proposed meta-analysis methods with two applied prediction model examples, one for diagnosis and one for prognosis, and compare the performance of different implementation strategies, including recalibration

3.1 Diagnostic example: Prediction of existing deep vein thrombosis

Data, model development and competing implementation strategies

We used IPD from 12 studies to develop a diagnostic prediction model for the risk of having deep vein thrombosis (DVT) in patients that were suspected of having DVT, as described previously [34]. A total of 10002 patients were available across the 12 studies (with study sample sizes ranging from 153 to 1768 patients), and 1864 (19%) patients truly had DVT. This IPD is used here only for illustration purposes and not to develop or recommend the optimal diagnostic model to be used in medical practice.

The prediction model was developed using logistic regression, including a separate intercept for each study and three predictors chosen *a priori*: *sex* (male=1, female=0), *surgery* (recent surgery or bedridden = 1, no recent surgery or bedridden = 0) and *calf difference* (≥3cm = 1, <3cm = 0). However, three different implementation strategies were considered (for the model intercept) when applying the developed model to the external validation dataset:

Strategy (1): Use a new intercept estimated in the external validation dataset itself. This is a form of model recalibration [35].

Strategy (2): Use the estimated weighted average of the study intercept terms from the developed model.

Strategy (3): Use the estimated intercept for one of the studies in the developed model that had the most similar prevalence of DVT to the external validation study.

Internal-external cross-validation was undertaken for each implementation strategy, and their predictive performance then summarised and compared across the 12 external validation studies using our multivariate meta-analysis approach.

Results

Regardless of which study was excluded, the predictor effect estimates (log odds ratios) were very similar in each cycle of the internal-external cross-validation approach (supplementary material 2 shows the parameter estimates in each cycle, and the intercept to be implemented in strategy (3)). During external validation of the model, for each implementation strategy four validation statistics were estimated: calibration-in-the-large, calibration slope, the C-statistic, and the ratio of expected and observed DVT cases, as defined in the Appendix. These results are shown (with standard errors) in supplementary material 3(a) for each of the strategies. Their within-study correlations, obtained from bootstrapping with 1000 samples, are shown in supplementary material 3(b). These are large (between +0.90 to +0.98) for the calibration slope and C- statistic, indicating a strong positive relationship between them. In other words, as the observed calibration slope of model predictions decreases (becomes flatter) the observed discrimination of the model predictions also decreases (less separation); conversely, when model predictions produce a steeper observed calibration slope the discrimination is improved. The other measures of calibration (calibration-in-the-large and expected/observed) measure overall agreement, and thus are not affected so much by changes in discrimination; thus their within-study correlation with the C-statistic is close to zero. There is a perfect negative correlation between log(expected/observed) and calibration-in-the-large by definition.

The multivariate meta-analysis results for each statistic are shown in Table 1. The meta-analysis results for the C-statistic are practically the same in all implementation strategies, as are those for the calibration slope. The mean C-statistic is 0.69 (95% CI: 0.67 to 0.71), indicating moderate discrimination. There is a small amount of between-study heterogeneity ($\hat{\tau}\approx0.02$; $I^2\approx37\%$), leading to a 95% prediction interval of 0.64 to 0.73, revealing fairly consistent discrimination performance across studies (Figure 1). The mean calibration slope is around 0.98 (95% CI: 0.85 to 1.10), which is close to the ideal value of one though indicating very slight over-prediction. The amount of between-study heterogeneity is large ($\hat{\tau}\approx0.16$; $I^2\approx59\%$), leading to a wide 95% prediction interval (e.g. 0.59 to 1.38 for strategy (2)). This contains values well above and well below one, which respectively suggest that in

some populations the predicted probabilities vary too little (i.e. the model is under-fitted and/or assigns probabilities that are too similar across individuals) and in others they vary too much (i.e. the model is over-fitted to the development sample and assigns probabilities that vary too much across individuals). This illustrates how the average performance is an incomplete picture; calibration slope is good on average, but could be poor in particular populations (Figure 2).

Calibration-in-the-large does differ more importantly across implementation strategies (Table 1), as it is sensitive to the choice of intercept. The meta-analysis results reveal it is, on average, slightly worse for strategy (1) as there is a small over-prediction in the proportion with DVT (-0.13, 95% CI: -0.19 to -0.08). However, there is almost no heterogeneity in the calibration-in-the-large ($\hat{\tau}$=0.008; I^2 =1%), leading to a narrow 95% prediction interval (-0.20 to -0.07). Using strategy (2) or (3) the average calibration-in-the-large is closer to zero (-0.004 and 0.047 respectively), but comes at the expense of slightly larger between-study heterogeneity ($\hat{\tau}$=0.53 and 0.27, I^2 =97% and 89% respectively), leading to wider prediction intervals. For example, for strategy (2) the 95% prediction interval is -1.24 to 1.23.

Instead of calibration-in-the-large, it is perhaps easier to interpret the Expected/Observed proportion of DVT cases (Table 1). This follows a similar pattern (Table 1), with narrowest prediction interval for strategy (1) and slightly improved average performance for strategies (2) and (3). The 95% prediction interval for Expected/Observed for strategy (1) suggests the overall agreement is likely to be reasonable in new populations (1.05 to 1.14), with the number of DVT cases over-predicted by between 5% and 14%. However, the 95% prediction interval is unsatisfactory for the other strategies; for example, it is 0.41 to 2.54 for strategy (2) indicating the number of predicted DVT cases in a new population could range from 59% too few up to 154% too many.

Overall, therefore, strategy (1) appears best as it removes heterogeneity in the calibration-in-the-large and Expected/Observed, whilst maintaining similar discrimination. However, the prediction model would benefit from additional predictors, as currently discrimination is only moderate and there is large heterogeneity in calibration slope. This is confirmed by a joint probability of only 0.03 that strategy (1) will give a C-statistic $\geq$ 0.7 and a calibration slope between 0.9 and 1.1 in a new population (Table 2). If the criteria for model discrimination is relaxed to a C-statistic $\geq$ 0.65, then the joint probability improves but only to 0.43.

3.2 Prognostic example: Prediction of mortality in breast cancer patients

Data, model development, and competing implementation strategies

We used IPD from eight cohort studies (relating to eight different countries from Look et al.[36]) to develop and evaluate a prognostic prediction model for the risk of mortality over time in women recently diagnosed with breast cancer. In total there were 7435 patients (ranging from 69 patients to 3242 per study) and 2043 events. The maximum follow-up duration was 120 months and the median follow-up duration across all studies was 86.3 months. Internal-external cross-validation was used and, in each cycle, a Royston-Parmar flexible parametric survival model was fitted [37-39], with the baseline cumulative hazard function modelled using restricted cubic splines (with four knots deemed sufficient) and predictor effects (hazard ratios) assumed constant over time. A set of eight candidate predictors was considered at each cycle: age, tumour type, tumour grade, tumour size, number of positive nodes, menopausal status, adjuvant therapy, and hormone receptor status. Backwards selection was used, with $p>0.05$ taken for exclusion. Separate but proportional baseline hazard functions were included for each country; that is, one study was taken as the reference group, and others were allowed a country-specific adjustment factor). When applying the developed model to the external validation study, three different implementation strategies were considered (in regard the baseline hazard):

Strategy (1): Use a new country-specific adjustment factor as estimated in the validation study itself. This is a form of recalibration, but assumes the baseline hazard in the validation study and the development studies are proportional.

Strategy (2): Use a weighted average of the estimated country-specific adjustment factors from the developed model.

Strategy (3): Use the country-specific adjustment factor for a country that was included in the developed model and is closest geographically to the validation country.

Internal-external cross-validation was undertaken for each strategy, and their predictive performance then summarised and compared across the eight external validation studies using meta-analysis.

Results

The predictor effect estimates (log hazard ratios) were similar in each cycle of the internal-external cross-validation approach (results available on request). The backwards selection retained all candidate predictors in each cycle, apart from menopausal status which was always excluded. For each implementation strategy, we evaluated model performance in each external validation study by estimating Harrell's C-statistic [20], the D-statistic [22, 40], and the calibration slope between the predicted hazard function and the observed hazard function, as defined in the Appendix. The estimates, with their variances and within-study correlation, are shown in supplementary material 4. Within-study correlations were all positive, and generally moderate to large.

Multivariate meta-analysis of the validation statistics is summarised in Table 3 for each implementation strategy. The summary C-statistic and D-statistic results are barely affected by the choice of strategy. The average C-statistic is 0.71 and its 95% prediction interval is 0.66 to 0.76, suggesting consistently moderate discrimination across populations. The average D-statistic is about 0.33, which equates to a moderate hazard ratio of 1.39 (95% CI: 1.23 to 1.57) between two equal sized groups across the prognostic index. However, D is inconsistent across populations (I^2 is about 87%), and thus its prediction interval is wide (Table 3).

Calibration slope is affected by the choice of strategy. For strategy (1), which allows recalibration in the validation study, the calibration slope is excellent. The meta-analysis gives an average calibration slope of 1.003, with only moderate heterogeneity (I^2 = 35%) leading to a narrow prediction interval of 0.93 to 1.08. In contrast, strategies (2) and (3) perform poorly. Although average calibration is excellent, there is large between-study heterogeneity (e.g. I^2 = 99% for strategy (3)) leading to wide predictions intervals (e.g. 0.15 to 1.77 for strategy (3)). This again reveals how average performance is an incomplete and potentially misleading summary of performance.

Figure 3 shows joint prediction ellipses for the C-statistic and the calibration slope, derived using the multivariate meta-analysis results for each strategy. For implementation strategy (1) there is a joint probability of 0.67 for a C-statistic $\geq$ 0.7 and a calibration slope between 0.9 and 1.1; however, the probability is only 0.15 for strategy (3) and 0.22 for strategy (2).

Strategy (1) thus performs best, but it requires recalibration of the model in new countries and may be difficult to implement. We therefore sought to improve strategy (2), which does not include recalibration, by identifying the cause of heterogeneity in its calibration performance. It was observed that Study 3 gave the poorest calibration slope upon external validation of the models (Table 4), most likely due to the baseline hazard in Study 3 being different in shape (non-proportional) to those other studies. Extending equation (3) to a multivariate meta-regression with a covariate for country (1= Study 3, 0 = otherwise) explained a large part of the heterogeneity ($p < 0.001$). We repeated the internal-external cross-validation approach for strategy (2), but omitted Study 3 for the entire process. External validation performance was improved, as heterogeneity in calibration slope was reduced ($\hat{\tau} = 0.156$ excluding Study 3, $\hat{\tau} = 0.22$ including Study 3), and thus its 95% prediction interval was narrower (supplementary material 5). The joint probability for a C-statistic≥0.7 and a calibration slope between 0.9 and 1.1 was improved to 0.32, but still considerably worse than strategy (1), indicating recalibration remains preferable.

4. Discussion

We have proposed a multivariate meta-analysis approach for summarising and comparing prediction model performance across multiple external validation studies using IPD. This can be used within internal-external cross-validation to also incorporate a model development phase, or when IPD from multiple studies are available for external validation of existing models. Each of the statistical methods involved (such as obtaining within-study correlations and fitting the multivariate equation) only take up to a few minutes to perform using computer software such as STATA, and provide results that improve the interrogation of a prediction model's performance and its implementation strategy.

Currently, most external validation research is undertaken using a single dataset. However, multivariate meta-analysis of IPD is a novel way to examine the overall performance and generalisability of a prediction model across multiple datasets[13, 15, 16]. A good model will have satisfactory performance on average across all external validation datasets. But ideally there should also be little or no between-study heterogeneity in performance. Our examples showed that a prediction model may have excellent *average* performance, but may not have *consistent* performance across datasets. Such heterogeneity is rarely considered in external validation research, but should be routinely examined where possible, in particular to identify

the best implementation strategy. In our examples, the investigation of heterogeneity revealed that recalibration of the intercept term to the validation population was essential, otherwise there was considerable inconsistency in calibration performance of our prediction models. The importance of intercept recalibration is also shown elsewhere [41, 42]. However, it may not entirely remove the issue of miscalibration, as seen in the DVT example where there remained slight over-prediction even after recalibration. In particular, if there is also heterogeneity in predictor effects then one may also need to recalibrate these to the intended population; however, this defeats the purpose of the initial research (i.e. to develop a prediction model that can be used widely and easily), and rather indicates that additional and/or more homogenous predictors are required.

Heterogeneity in discrimination performance was also observed in our examples. This may also be due to heterogeneity in predictor effects across populations and/or different case-mix distributions across populations, as populations with wider ranges of continuous predictors often have better discrimination[15]. For such reasons, incorporating matched case-control studies alongside cohort studies may increase heterogeneity in discrimination performance, as the former typically have narrower ranges of predictors [43]. Another potential cause of heterogeneity in performance of a prognostic prediction model is follow-up time, and also heavy censoring may bias Harrell's C-statistic, prompting Gönen and Heller to propose an alternative [44]. Such factors may also impact the magnitude of between-study correlation in the performance measures.

As external validation of a prediction model usually requires multiple statistical measures of performance, in particular at least one for calibration and one for discrimination [8, 19], our multivariate meta-analysis approach jointly synthesises all measures together across multiple validation studies. This accounts for their within-study and between-study correlation [45], which may arise because measures are highly related [33]. For example, the C-statistic and D-statistic typically have moderate to large positive within-study correlation (as seen in supplementary material 4(b)) as they are both measures of discrimination and within-studies are estimated on the same patients. Similarly the calibration slope and C-statistic may also be correlated between-studies, for instance if the between-study heterogeneity in predictor effects causes calibration slope to become greater than 1 as discrimination improves, but less than 1 as discrimination worsens. Accounting for such correlation in the meta-analysis allows the borrowing of strength across performance measures to potentially reduce bias and

improve precision [46, 47]. Further, it is crucial to account for correlation when computing joint probabilities of model performance, such as the magnitude of the C-statistic and calibration slope, as otherwise inferences may be misleading [45].

Our intention was to illustrate how the multivariate meta-analysis approach allows researchers to summarise both discrimination and calibration. We focused on well-known statistical criteria, such as the calibration slope, (Harrell's) C-statistic, and Royston and Sauerbrei's D-statistic. However, we recognise that the criteria for a 'good' prediction model is open to much debate [48], and readers may prefer to meta-analyse other statistical measures available, including alternatives to Harrell's C-statistic [44]. Clinical criteria may also be preferred [49], to focus more on the consequences for decision-making [50].Whatever criteria are used, we recommend they are pre-specified in a published protocol [51]. Visual plots of calibration [23] and discrimination [52] are also important, as neatly illustrated by Royston et al.[17]. Calibration estimates can also be obtained (and then meta-analysed) for particular subgroups within studies, for example defined by particular patient characteristics or categories of the prognostic index [23]. Also, we note that excellent validation performance is not the end of the story: a prediction model's impact on patient outcomes also needs to be evaluated, for example in subsequent trials[3].

A potential limitation of our work is the multivariate normality assumption for the distribution of true performance across studies. Though this is a common assumption in the meta-analysis field, prediction intervals and regions are potentially vulnerable to departures from this[53]. A related issue is the choice of scale to use for the estimates of validation performance[16], and further research is needed on this. Internal-external cross-validation is also limited if the number of studies are small, and researchers should ensure the number of events is suitable in each cycle [54, 55].

In conclusion, we propose multivariate meta-analysis for external validation of the performance and implementation of a prediction model when IPD are available for multiple studies. The approach encourages researchers to focus not only on average performance, but also on the consistency in performance across populations, for both calibration and discrimination.

References

[1] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.

[2] Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. British Medical Journal. 2009;338:b604 (1373-7).

[3] 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:e1001381.

[4] Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ. 2007;335:136.

[5] Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Res Treat. 1992;22:207-19.

[6] 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:e1001380.

[7] Steyerberg EW, Harrell FEJ, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774-81.

[8] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605.

[9] Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in, prediction research: A clinical example. Journal of Clinical Epidemiology. 2003;56:826-32.

[10] Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. BMJ. 2012;344:e4181.

[11] Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015; 68:279-89.

[12] Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. BMC medicine. 2010;8:21.

[13] Ahmed I, Debray TP, Moons KG, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Med Res Methodol. 2014;14:3.

[14] Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ. 2010;340:c221.

[15] Pennells L, Kaptoge S, White IR, Thompson SG, Wood AM. Assessing risk prediction models using individual participant data from multiple studies. Am J Epidemiol. 2014;179:621-32.

[16] van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. BMC Med Res Methodol. 2014;14:5.

[17] Royston P, Parmar MKB, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. Statistics in Medicine 2004;23:907-26.

[18] Debray TP, Moons KG, 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;32:3158-8.

[19] Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40.

[20] Harrell FE. Regression modeling strategies, with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.

[21] Pencina MJ, D'Agostino RB, Song L. Quantifying discrimination of Framingham risk functions with different survival C statistics. Stat Med. 2012;31:1543-53.

[22] Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med. 2004;23:723-48.

[23] Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 2013;13:33.

[24] Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society, Series A. 2009;172:137-59.

[25] Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011;342:d549.

[26] White IR. Multivariate random-effects meta-regression: Updates to mvmeta. The STATA Journal 2011;11:255-70.

[27] White IR. Multivariate meta-analysis. The Stata Journal. 2009;9:40-56.

[28] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Bmj. 2003;327:557-60.

[29] Rucker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol. 2008;8:79.

[30] Jackson D, Riley RD, White IR. Multivariate meta-analysis: potential and promise. Stat Med. 2011;30:2481-98.

[31] Jackson D, White IR, Riley RD. A matrix-based method of moments for fitting the multivariate random effects model for meta-analysis and meta-regression. Biom J. 2013;55:231-45.

[32] Azzalini A, Genz A. The R package 'mnormt': The multivariate normal and 't' distributions (version 1.5-1).URL: http://azzalini.stat.unipd.it/SW/Pkg-mnormt.

[33] Riley RD, Price MJ, Jackson D, Wardle M, Gueyffier F, Wang J, et al. Multivariate meta-analysis using individual participant data. Res Synth Method. 2015 (in-press).

[34] Geersing GJ, Zuithoff NP, Kearon C, Anderson DR, Ten Cate-Hoek AJ, Elf JL, et al. Exclusion of deep vein thrombosis using the Wells rule in clinically important subgroups: individual patient data meta-analysis. BMJ. 2014;348:g1340.

[35] Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med. 2004;23:2567-86.

[36] Look MP, van Putten WL, Duffy MJ, Harbeck N, Christensen IJ, Thomssen C, et al. Pooled analysis of prognostic impact of urokinase-type plasminogen activator and its inhibitor PAI-1 in 8377 breast cancer patients. J Natl Cancer Inst. 2002;94:116-28.

[37] Royston P. Flexible parametric alternatives to the Cox model, and more. The Stata Journal. 2001;1:1-28.

[38] Lambert PC, Royston P. Further developments of flexible parametric models for survival analysis. The Stata Journal 2009;9:265-90.

[39] Royston P, Lambert PC. Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model: CRC Press; 2011.

[40] Royston P. Explained variation for survival models. Stata Journal 2006;6:83-96.

[41] Schuetz P, Koller M, Christ-Crain M, Steyerberg E, Stolz D, Muller C, et al. Predicting mortality with pneumonia severity scores: importance of model recalibration to local settings. Epidemiol Infect. 2008;136:1628-37.

[42] Janssen KJ, Moons KG, 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.

[43] Janes H, Pepe MS. Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value Biometrics. 2008;64:1-9.

[44] Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. Biometrika. 2005;92:965-70.

[45] Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. JRSS Series A. 2009;172(4):789-811.

[46] Kirkham JJ, Riley RD, Williamson PR. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. Stat Med. 2012;31:2179-95.

[47] Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. Stat Med. 2007;26:78-97.

[48] Vach W. Calibration of clinical prediction rules does not just assess bias. J Clin Epidemiol. 2013;66:1296-301.

[49] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19:453-73.

[50] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565-74.

[51] Peat G, Riley RD, Croft P, Morley KI, Kyzas PA, Moons KG, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. PLoS Med. 2014;11:e1001671.

[52] Royston P, Altman DG. Visualizing and assessing discrimination in the logistic regression model. Stat Med. 2010;29:2508-20.

[53] Lee KJ, Thompson SG. Flexible parametric models for random-effects distributions. Stat Med. 2008;27:418-34.

[54] Peduzzi PN, Concato J, Kemper E, Holford T, Feinstein A. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49:1373-9.

[55] Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995;48:1503-10.

Table 1: Trivariate meta-analysis results* for the calibration and discrimination performance of the DVT model for each implementation strategy.

Strategy	Validation statistic	Estimate (95% CI) of mean, μ	95% prediction interval	I^2 %	$\hat{\tau}$ (95% CI)
Strategy (1): Develop using logistic regression and implement with intercept estimated in external validation study	Calibration-in-the-large	-0.130 (-0.185 to -0.075)	-0.195 to -0.065	1	0.008
	Calibration slope	0.975 (0.855 to 1.097)	0.597 to 1.353	57	0.158
	Log(Expected/Observed)	0.086 (0.047 to 0.124)	0.041 to 0.128	0	0.0009
	C-statistic	0.687 (0.670 to 0.704)	0.645 to 0.729	34	0.017
Strategy (2): Develop using logistic regression and implement with average study intercept taken from developed model	Calibration-in-the-large	-0.004 (-0.313 to 0.305)	-1.240 to 1.232	97	0.532
	Calibration slope	0.980 (0.853 to 1.107)	0.585 to 1.375	59	0.165
	Log(Expected/Observed)	0.022 (-0.206 to 0.250)	-0.887 to 0.931	97	0.391
	C-statistic	0.687 (0.669 to 0.705)	0.640 to 0.734	37	0.019
Strategy (3): Develop using logistic regression and implement with intercept taken from a study used in development data with a similar prevalence	Calibration-in-the-large	0.047 (-0.120 to 0.214)	-0.584 to 0.678	89	0.270
	Calibration slope	0.976 (0.851 to 1.102)	0.578 to 1.375	59	0.167
	Log(Expected/Observed)	-0.029 (-0.150 to 0.093)	-0.485 to 0.427	89	0.195
	C-statistic	0.687 (0.669 to 0.705)	0.640 to 0.734	38	0.019

* A trivariate meta-analysis was fitted to calibration-in-the-large, calibration slope and C-statistic, and then again for log(Expected/Observed), calibration slope, and C-statistic. Perfect negative correlation between calibration-in-the-large and expected/observed within studies prevents all four measures being analysed together (due to collinearity). Results were practically the same for calibration slope and C-statistic, regardless of the trivariate model fitted.

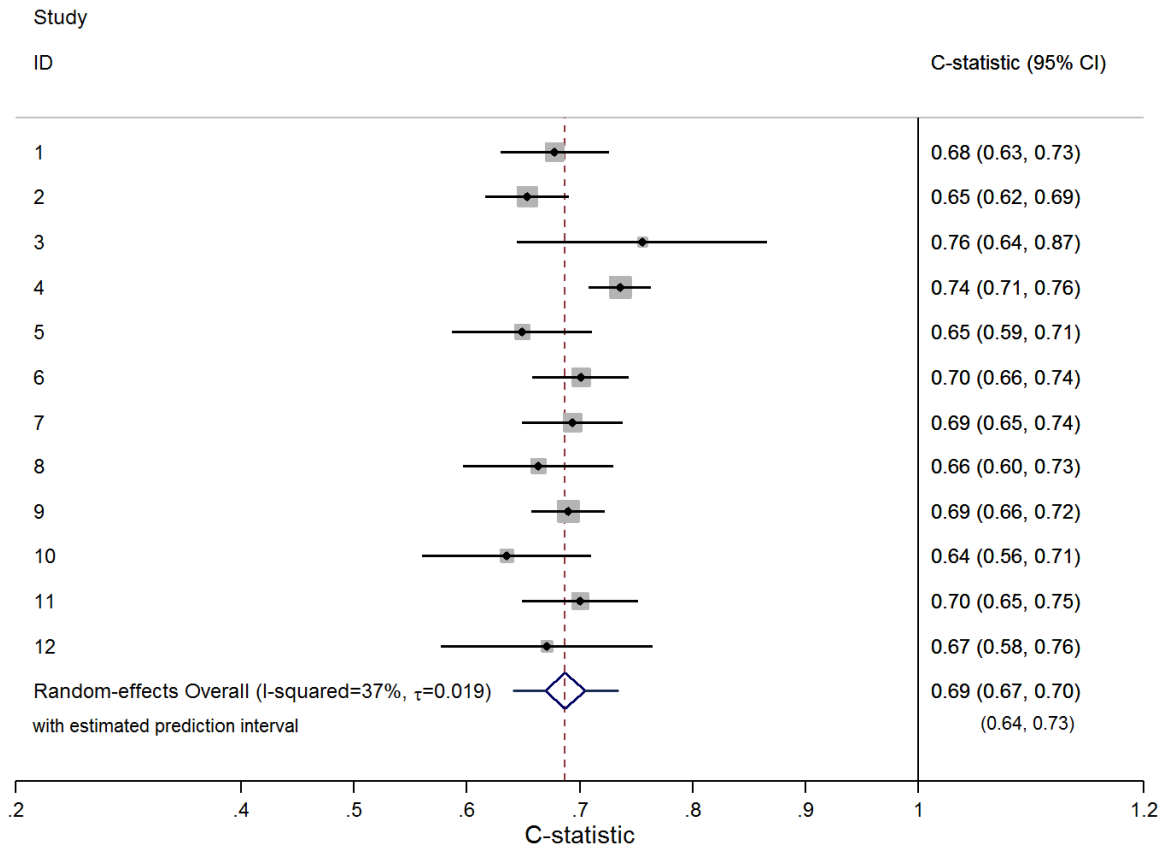
Table 2: Joint predicted probability of 'good' discrimination and calibration performance for each of the three DVT models, derived using the multivariate meta-analysis results for the C-statistic and calibration slope shown in Table 1.

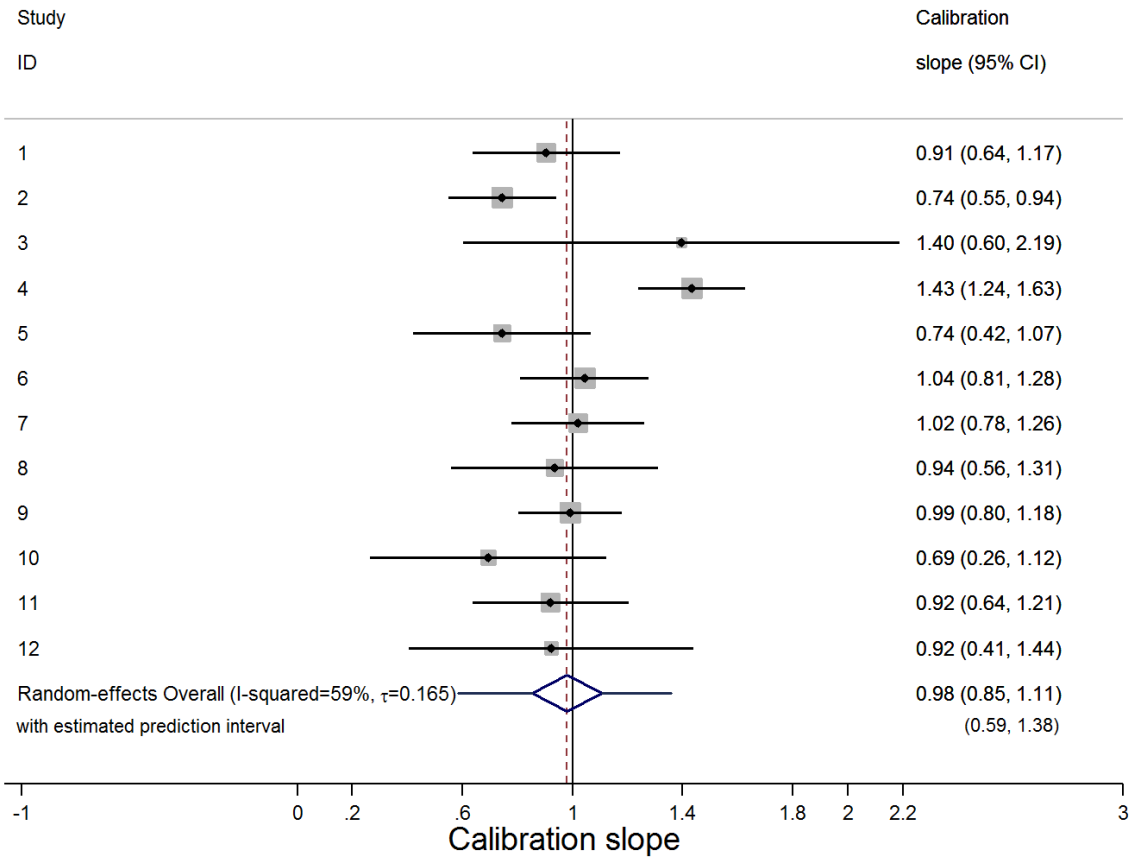
Calibration slope required	**Minimum C-statistic required**	**Joint predicted probability of meeting criteria in new population**		
		Strategy (1): Develop using logistic regression and implement with intercept estimated in external validation study	**Strategy (2):** Develop using logistic regression and implement with average study intercept taken from developed model	**Strategy (3):** Develop using logistic regression and implement with intercept taken from a study used in development data with a similar prevalence
0.9 to 1.1	0.70	0.027	0.037	0.037
0.8 to 1.2	0.70	0.146	0.158	0.156
0.9 to 1.1	0.65	0.427	0.413	0.409
0.8 to 1.2	0.65	0.728	0.712	0.707

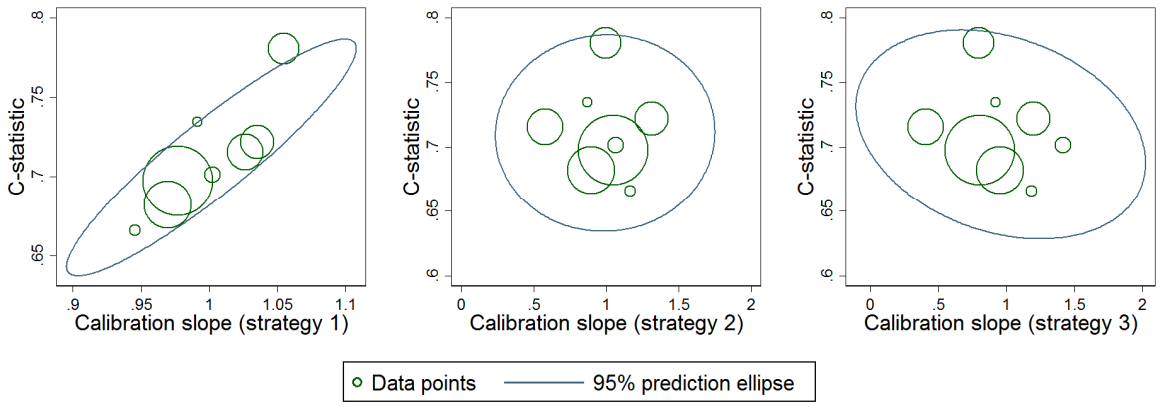
Table 3: Trivariate random-effects meta-analysis results for calibration and discrimination performance of the breast cancer model for each implementation strategy

Strategy	Validation statistic	Pooled estimate (95% CI)	95% prediction interval	I-squared	Estimate of τ	Joint probability of 'good'* performance in a new population
Strategy (1): Develop using Royston-Parmar model and implement with baseline hazard estimated in validation study	Calibration slope	1.003 (0.971 to 1.036)	0.927 to 1.080	35%	0.026	0.67
	C-statistic	0.711 (0.690 to 0.733)	0.657 to 0.766	49%	0.019	
	D-statistic	0.328 (0.215 to 0.442)	-0.056 to 0.713	87%	0.146	
Strategy (2): Develop using Royston-Parmar model and implement with the estimated average baseline hazard from developed model	Calibration slope	0.994 (0.835 to 1.153)	0.411 to 1.577	98%	0.224	0.22
	C-statistic	0.711 (0.691 to 0.732)	0.662 to 0.761	43%	0.017	
	D-statistic	0.332 (0.212 to 0.452)	-0.080 to 0.745	88%	0.157	
Strategy (3): Develop using Royston-Parmar model and implement with the estimated baseline hazard from the closest geographical country	Calibration slope	0.961 (0.741 to 1.181)	0.148 to 1.775	99%	0.313	0.15
	C-statistic	0.710 (0.687 to 0.734)	0.653 to 0.767	50%	0.020	
	D-statistic	0.330 (0.211 to 0.450)	-0.068 to 0.728	87%	0.151	

* defined by a C-statistic≥0.7 and an calibration slope between 0.9 and 1.1







Appendix: Brief explanation of key statistical concepts considered in the article

C-statistic

A measure of a prediction model's discrimination (separation) between those with and without an event (outcome), which can be calculated for either binary or survival outcome data [20, 21]. Also known as the concordance index or, for binary outcomes, the area under the receiver operating characteristic (ROC) curve. It gives the probability that for any randomly selected pair of individuals, one with and one without the event (outcome), the model assigns a higher probability to the individual with the event (outcome). A value of 1 indicates the model has perfect discrimination, whilst a value of 0.5 indicates the model discriminates no better than chance. For the DVT model we used the 'roctab' module in STATA to calculate the C-statistic, whilst for the breast cancer model we used 'stcstat2' to calculate Harrell's C-statistic.

D-statistic

A measure of discrimination for time-to-event outcomes [22]. This can be interpreted as the log hazard ratio comparing two equally sized groups defined by dichotomising at the median value of the prognostic index from the developed model (where the prognostic index is defined by the combined predictor effects in the developed model, i.e. beta1*X1 + beta2*X2 + ...). Higher values for the D-statistic indicate greater discrimination, and an increase of 0.1 over other risk scores is suggested to be a good indicator of improved prognostic separation [10]. For the breast cancer model, the D-statistic in the external validation study was calculated using the 'str2d' module [40] in STATA after fitting the Royston-Parmar model in our development data using the 'stpm2' module [38].

Calibration slope

This is a measure of agreement between observed and predicted risk of the event (outcome) across the whole range of predicted values [1, 20]. For example, if a prediction model is developed using logistic regression it is of the form:

$$\text{logit}(p) = \text{alpha} + \text{beta1}*X1 + \text{beta2}*X2 + ...$$

Then the predicted probability (P-pred) is derived using

$$\text{P-pred} = \exp(\text{alpha} + \text{beta1}*X1 + \text{beta2}*X2 + ...) / [1 + \exp(\text{alpha} + \text{beta1}*X1 + \text{beta2}*X2 + ...)]$$

In the validation data, fitting the model

$$\text{logit}(p) = \text{delta0} + \text{delta1}*(\text{logit}(\text{P-pred}))$$

is used to calculate the calibration slope (delta1), which should ideally be 1 (though a value of 1 does not by itself confirm calibration is perfect [48]).

In the breast cancer example in our article, we used a flexible parametric survival model to develop the prediction model, using the Royston-Parmar approach where the log cumulative hazard function is modelled using restricted cubic splines [37-39], with four knots chosen. To estimate calibration slope of the developed model, we fitted the following model in each external validation study:

$$\text{lnH}(t) = \text{gamma0}+\text{gamma1}*z1+\text{gamma2}*z2+\text{gamma3}*z3 +b*X$$

where b is the calibration slope, lnH(t) is the log cumulative hazard function over time, t, and the gamma and z terms define the knots and the baseline lnH(t). The (gamma1*z1+gamma2*z2+gamma3*z3) value was forced to be that from the developed model; the X value for each patient corresponds to their value of the prognostic index from the developed model (i.e. Beta1*X1 + Beta2*X2 + ...); and the gamma0 value is dependent on the implementation strategy used. In strategy (1) gamma0 is newly estimated in the external validation study (akin to recalibration). In strategy (2) gamma0 is forced to be a weighted (meta-analysis) average of all the study-specific gamma0 estimates from the developed model. In strategy (3) the gamma0 value is forced to be the gamma0 estimate for the nearest country available in the developed model.

Calibration-in-the-large

This is an overall measure of calibration,[1] defined by fitting in the validation dataset

$$\text{logit}(p) = \text{alpha0} + \text{offset}$$

where alpha0 is the calibration-in-the-large and the offset term is equal to logit(P-pred)

Expected/Observed number of events (E/O)

This is another measure of calibration, closely related to the calibration-in-the-large, but more intuitive to interpret. It provides the ratio of the total expected events (outcomes) to the total observed events (outcomes). It can be obtained by summing all predicted probabilities in the

validation dataset and dividing by the number of observed events. An ideal value is 1. Values less than 1 indicate the model is under-predicting the total number of events in the population, whilst values above 1 indicate it is over-predicting the total events in the population.

Non-parametric bootstrapping to obtain variance-covariance matrix of performance estimates

Consider a single external validation study. Bootstrapping uses the IPD from this study and randomly selects one patient with replacement, then randomly selects a second patient with replacement, and repeats until the same sample size is obtained as in the original study. This process is repeated *b* times, so that *b* bootstrap samples are obtained. Then, in each of the bootstrap samples the performance statistics of the developed model are estimated (e.g. Y_{i1} could be the estimated C-statistic and Y_{i2} could be the estimated calibration slope estimate in study *i*). This produces *b* values for each performance statistic. The observed variance of the *b* values for each statistic estimates their variance (i.e. it gives S_{i1}^2, the variance of Y_{i1} , etc). Similarly, the observed correlation across the *b* samples for each pair of validation statistics estimates their within-study correlation (i.e. it gives $\rho_{Wi(1,2)}$, the within-study correlation between Y_{i1} and Y_{i2} etc).

Bivariate model

When there are two performance measures of interest (e.g. C-statistic and calibration slope), the general multivariate meta-analysis of equation (3) can be simplified to the following bivariate meta-analysis:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix}, \mathbf{S_i} \right) \qquad \mathbf{S_i} = \begin{pmatrix} S_{i1}^2 & \rho_{Wi(1,2)} S_{i1} S_{i2} \\ \rho_{Wi(1,2)} S_{i1} S_{i2} & S_{i2}^2 \end{pmatrix}$$

$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \mathbf{\Sigma} \right) \qquad \mathbf{\Sigma} = \begin{pmatrix} \tau_1^2 & \rho_{B(1,2)} \tau_1 \tau_2 \\ \rho_{B(1,2)} \tau_1 \tau_2 & \tau_2^2 \end{pmatrix}$$

Supplementary material 1: Example of SAS syntax to derive joint probabilities of model performance with regard to two performance measures.

The joint probability of 'good' performance in a new population is obtained below by calculating the proportion of 1000000 samples drawn from the approximate posterior bivariate t-distribution (with *k* - 2 d.f.; k = no. of studies). The parameters values to use are derived from the estimates of the mean vector and its variance-covariance matrix, and the between-study variance-covariance matrix, following REML estimation of multivariate meta-analysis equation (3). SAS code is provided below based on using the 'RANDMVT' module . Alternatively this can be performed using the "mnormt" package in R software [32].

```
/* Breast cancer example: strategy (1)*/
proc iml;
load module=randmvt;
/* set random number seed */
call randseed(1);
/* define number of samples to take */
N=100000;
/* enter degrees of freedom (no. of studies minus 2)*/
DF = 6;
/* Define the means of the posterior distribution to be the meta-analysis mean
estimates (mu) for the C-statistic and calibration slope*/
Mean = {0.711 1.003};
/* Define the variance-covariance matrix S for the posterior distribution by:
V_11 = Variance for C-statistic = tau1-squared + var(mu1)
V_22 = Variance for calibration slope = tau2-squared + var(mu2)
V_12 = V_21 = Covariance of the C-statistic and calibration slope  = tau12 +
cov(mu1,mu2)
(where tau1 and tau2 are the between-study SD for the C-statistic and the
calibration slope, respectively; mu1 and mu2 are the mean estimates of the C-
statistic and the calibration slope, respectively; var(mu1) and var(mu2) are the
variances of the estimates mu1 and mu2, respectively; tau12 is the between-study
covariance; and cov(mu1,mu2) is the covariance of the estimates of mu1 and mu2) */
S = {0.00048 0.000611, 0.000611 0.000958};
/* derive the samples and define them by x*/
x = RANDMVT( N, DF, Mean, S );
/* show the 1000 pair of values generated if on interest*/
print x;

/* convert x to a dataset called samples */
create samples from x;
append from x;
quit;

/* Create a second dataset that identifies whether 'good' performance was obtained
for each sample within x - here 'good' is defined by the C-statistic > 0.7 and the
calibration slope between 0.9 and 1,1 */
data samples2;
set samples;
/* C-statistic criteria */
if col1 > 0.7 then y =1;
/* calibration criteria */
if col2 > 0.9 then z1 = 1;
if col2 < 1.1 then z2 = 1;
/* joint criteria */
if y=1 and z1=1 and z2 =1 then accept=1;
else accept=0;
run;

/* Calculate the joint probability of good performance, simply by the mean value of
the 'accept' variable - which is the proportion of samples that had the 'good'
performance */
proc means;
vars accept;
run;
```

Supplementary material 2: Model parameter estimates for the fitted DVT model to be implemented using strategy (3), which fits a logistic regression model in each cycle of the internal-external cross-validation approach to obtain predictor effects and study-specific intercepts.

Study used for external validation	**Study-specific intercept (a_i)**												**Age ($\hat{\beta}_1$)**	**Sex ($\hat{\beta}_2$)**	**Calf ($\hat{\beta}_3$)**
	Study 1	Study 2	Study 3	Study 4	Study 5	Study 6	Study 7	Study 8	Study 9	Study 10	Study 11	Study 12			
Study 1	-	-1.256	-2.528	-1.807	-2.511	-2.316	-3.061	-1.839	-2.165	-2.339	-2.038	**-2.788**	0.372	0.606	1.304
Study 2	-2.694	-	-2.555	-1.823	-2.538	-2.338	-3.080	**-1.857**	-2.192	-2.353	-2.059	-2.817	0.370	0.624	1.344
Study 3	-2.670	-1.253	-	-1.805	**-2.504**	-2.312	-3.062	-1.840	-2.157	-2.336	-2.033	-2.779	0.400	0.556	1.286
Study 4	-2.604	-1.191	-2.443	-	-2.426	-2.245	-3.005	**-1.785**	-2.082	-2.289	-1.967	-2.700	0.366	0.513	1.197
Study 5	-2.682	-1.265	**-2.538**	-1.814	-	-2.324	-3.070	-1.849	-2.173	-2.343	-2.045	-2.797	0.387	0.579	1.315
Study 6	-2.672	-1.255	-2.529	-1.811	**-2.508**	-	-3.066	-1.843	-2.161	-2.343	-2.041	-2.783	0.409	0.589	1.277
Study 7	**-2.672**	-1.255	-2.529	-1.810	-2.509	-2.319	-	-1.842	-2.162	-2.342	-2.040	-2.784	0.401	0.594	1.283
Study 8	-2.677	-1.260	-2.532	**-1.812**	-2.513	-2.321	-3.068	-	-2.166	-2.342	-2.042	-2.788	0.403	0.572	1.294
Study 9	-2.660	-1.245	-2.515	-1.797	-2.498	-2.304	-3.050	-1.828	-	-2.329	**-2.026**	-2.777	0.350	0.611	1.300
Study 10	-2.677	-1.260	-2.531	-1.809	-2.515	-2.318	-3.066	-1.844	-2.168	-	-2.039	**-2.791**	0.381	0.572	1.313
Study 11	-2.678	-1.261	-2.533	-1.812	-2.515	-2.321	-3.069	-1.847	**-2.167**	-2.342	-	-2.790	0.399	0.565	1.301
Study 12	-2.672	-1.256	-2.528	-1.808	-2.509	-2.317	-3.064	-1.841	-2.162	**-2.339**	-2.038	-	0.391	0.585	1.293

Note: Bold numbers represent the intercept used for external validation in the excluded study for strategy (3), where the intercept from the study with the closest prevalence was selected.

Supplementary material 3(a): Estimates (standard errors) of the calibration and discrimination performance of the DVT model developed in each cycle of the internal-external cross-validation approach, for the three implementation strategies

Study used for external validation	**Strategy (1):** Develop using logistic regression and apply with intercept estimated in external validation study				**Strategy (2):** Develop using logistic regression and apply with average study intercept taken from developed model				**Strategy (3):** Develop using logistic regression and apply with intercept taken from a study used in development data with a similar prevalence			
	CITL	**Calibration slope**	**Log(E/O)**	**C-statistic**	**CITL**	**Calibration slope**	**Log(E/O)**	**C-statistic**	**CITL**	**Calibration slope**	**Log(E/O)**	**C-statistic**
1	-0.172	0.903	0.140	0.678	-0.440	0.905	0.349	0.678	0.114	0.905	-0.094	0.678
	(0.098)	(0.131)	(0.080)	(0.024)	(0.098)	(0.136)	(0.081)	(0.025)	(0.094)	(0.132)	(0.078)	(0.024)
2	-0.051	0.741	0.028	0.653	1.105	0.745	-0.709	0.653	0.583	0.736	-0.344	0.652
	(0.079)	(0.100)	(0.042)	(0.019)	(0.078)	(0.100)	(0.043)	(0.019)	(0.084)	(0.102)	(0.045)	(0.019)
3	-0.172	1.418	0.135	0.761	-0.292	1.396	0.223	0.756	-0.042	1.390	0.036	0.755
	(0.224)	(0.397)	(0.174)	(0.055)	(0.225)	(0.405)	(0.176)	(0.057)	(0.223)	(0.408)	(0.173)	(0.057)
4	-0.084	1.432	0.060	0.735	0.488	1.434	-0.367	0.736	0.031	1.434	-0.022	0.736
	(0.054)	(0.100)	(0.039)	(0.014)	(0.057)	(0.099)	(0.040)	(0.014)	(0.057)	(0.102)	(0.040)	(0.014)
5	-0.185	0.742	0.141	0.649	-0.267	0.744	0.201	0.649	0.016	0.749	-0.011	0.650
	(0.122)	(0.164)	(0.094)	(0.031)	(0.125)	(0.165)	(0.096)	(0.032)	(0.121)	(0.159)	(0.093)	(0.031)
6	-0.149	1.030	0.112	0.699	-0.055	1.044	0.042	0.701	0.187	1.035	-0.144	0.700
	(0.082)	(0.114)	(0.062)	(0.021)	(0.084)	(0.119)	(0.064)	(0.022)	(0.084)	(0.119)	(0.063)	(0.022)
7	-0.178	1.017	0.156	0.694	-0.877	1.020	0.732	0.694	-0.399	1.014	0.344	0.693
	(0.090)	(0.117)	(0.080)	(0.023)	(0.089)	(0.122)	(0.078)	(0.023)	(0.090)	(0.123)	(0.080)	(0.024)
8	-0.115	0.932	0.081	0.663	0.464	0.936	-0.340	0.663	-0.038	0.943	0.028	0.665
	(0.133)	(0.189)	(0.093)	(0.035)	(0.129)	(0.192)	(0.090)	(0.034)	(0.132)	(0.192)	(0.092)	(0.034)
9	-0.139	0.994	0.098	0.690	0.118	0.991	-0.084	0.689	-0.132	0.996	0.093	0.690
	(0.068)	(0.099)	(0.048)	(0.017)	(0.073)	(0.096)	(0.052)	(0.017)	(0.070)	(0.104)	(0.050)	(0.017)
10	-0.127	0.695	0.103	0.636	-0.081	0.693	0.066	0.635	0.447	0.699	-0.373	0.637
	(0.150)	(0.215)	(0.122)	(0.037)	(0.144)	(0.219)	(0.116)	(0.038)	(0.146)	(0.208)	(0.119)	(0.036)
11	-0.135	0.921	0.094	0.701	0.258	0.921	-0.185	0.700	0.132	0.916	-0.093	0.700
	(0.111)	(0.140)	(0.078)	(0.026)	(0.111)	(0.145)	(0.078)	(0.026)	(0.110)	(0.140)	(0.077)	(0.026)
12	-0.197	0.936	0.160	0.673	-0.570	0.923	0.440	0.671	-0.458	0.930	0.359	0.672
	(0.191)	(0.269)	(0.155)	(0.048)	(0.190)	(0.264)	(0.155)	(0.048)	(0.193)	(0.256)	(0.157)	(0.046)

Note: CITL refers to calibration-in-the-large and log(E/O) refers to log of the Expected/Observed number of events.

Supplementary material 3(b): Within-study correlations (ρ_{Wi}), obtained through bootstrapping, between performance statistics estimated for the DVT model in each cycle of the internal-external cross-validation approach, for the three implementation strategies

	Study	CITL & calibration slope	CITL & log(E/O)	CITL & C-statistic	Calibration slope & log(E/O)	Calibration slope & C-statistic	Log(E/O) & C-statistic
Strategy (1): Develop using logistic regression and apply with intercept estimated in external validation study	1	-0.006	-1.000	-0.022	0.006	0.961	0.021
	2	-0.032	-1.000	-0.046	0.033	0.977	0.046
	3	-0.001	-0.999	0.009	0.001	0.955	-0.011
	4	0.118	-1.000	0.045	-0.117	0.919	-0.045
	5	0.046	-1.000	0.029	-0.046	0.983	-0.029
	6	-0.010	-1.000	-0.045	0.011	0.948	0.043
	7	0.047	-1.000	0.002	-0.047	0.912	-0.003
	8	0.071	-1.000	0.032	-0.072	0.953	-0.034
	9	-0.005	-1.000	-0.011	0.005	0.980	0.011
	10	0.108	-1.000	0.064	-0.108	0.900	-0.064
	11	0.000	-1.000	-0.025	0.002	0.956	0.026
	12	-0.035	-1.000	-0.029	0.036	0.980	0.030
Strategy (2): Develop using logistic regression and apply with average study intercept taken from developed model	1	-0.051	-1.000	-0.054	0.053	0.960	0.054
	2	0.046	-0.990	0.037	-0.054	0.976	-0.035
	3	-0.051	-0.999	-0.037	0.052	0.958	0.037
	4	0.062	-0.999	-0.009	-0.065	0.928	0.014
	5	0.069	-1.000	0.055	-0.069	0.981	-0.056
	6	0.031	-1.000	-0.013	-0.031	0.959	0.013
	7	0.018	-0.999	-0.025	-0.013	0.923	0.025
	8	0.105	-0.999	0.042	-0.106	0.951	-0.038
	9	0.074	-1.000	0.068	-0.074	0.976	-0.068
	10	0.035	-1.000	0.012	-0.034	0.895	-0.012
	11	0.053	-1.000	0.015	-0.052	0.953	-0.012
	12	0.000	-0.999	-0.001	-0.003	0.981	-0.003
Strategy (3): Develop using logistic regression and apply with intercept taken from a study used in development data with a similar prevalence	1	0.005	-1.000	-0.012	-0.004	0.963	0.013
	2	-0.006	-0.997	-0.033	0.002	0.978	0.034
	3	-0.053	-0.999	-0.049	0.054	0.959	0.050
	4	0.082	-1.000	0.015	-0.082	0.925	-0.014
	5	0.034	-1.000	0.022	-0.033	0.982	-0.022
	6	0.063	-1.000	0.020	-0.065	0.956	-0.019
	7	-0.014	-1.000	-0.025	0.015	0.925	0.025
	8	-0.002	-1.000	-0.052	0.001	0.954	0.051
	9	-0.010	-1.000	-0.021	0.011	0.981	0.021
	10	0.020	-0.999	-0.043	-0.024	0.892	0.043
	11	0.038	-1.000	-0.001	-0.040	0.956	0.000
	12	-0.036	-0.999	-0.041	0.036	0.980	0.039

Note: CITL refers to calibration-in-the-large and log(E/O) refers to log of the Expected/Observed number of events.

Supplementary material 4(a): Estimates (standard errors) of calibration and discrimination performance for the breast cancer model in each cycle of the internal-external cross-validation approach, for the three implementations strategies

Study excluded for external validation	C-statistic[1]	D-statistic[1]	Calibration slope[2]		
			Strategy (1): Develop using Royston- Parmar model and implement with baseline hazard estimated in validation study	**Strategy (2):** Develop using Royston-Parmar model and implement with the estimated average baseline hazard in develop model	**Strategy (3):** Develop using Royston- Parmar model and implement with the estimated baseline hazard from the closest geographical country
1	0.697 (0.008)	0.493 (0.027)	0.977 (0.012)	1.049 (0.012)	0.805 (0.012)
2	0.701 (0.036)	0.420 (0.117)	1.002 (0.057)	1.066 (0.057)	1.414 (0.056)
3	0.715 (0.023)	0.106 (0.056)	1.026 (0.036)	0.578 (0.037)	0.405 (0.037)
4	0.735 (0.068)	0.326 (0.187)	0.991 (0.097)	0.870 (0.098)	0.919 (0.097)
5	0.666 (0.050)	0.238 (0.168)	0.946 (0.088)	1.168 (0.086)	1.184 (0.086)
6	0.682 (0.017)	0.182 (0.041)	0.969 (0.037)	0.896 (0.038)	0.951 (0.037)
7	0.781 (0.027)	0.280 (0.063)	1.054 (0.052)	0.996 (0.053)	0.794 (0.054)
8	0.722 (0.016)	0.541 (0.058)	1.035 (0.030)	1.315 (0.029)	1.197 (0.030)

[1] The C-statistic and D-statistic only depend on the prognostic index (see Appendix). As the prognostic index (beta terms) from the developed model is not dependant on the implementation strategy, the C-statistic and D-statistic estimates are identical regardless of the implementation strategy used.

[2] obtained as defined in the Appendix.

Supplementary material 4(b): Within-study correlations (ρ_{Wi}), obtained using bootstrapping, between performance statistics estimated for the breast cancer model in each internal-external cross-validation cycle, for each implementation strategy

Country	C-statistic & D-statistic	C-statistic & Calibration slope (strategy (1))	C-statistic & Calibration slope (strategy (2))	C-statistic & Calibration slope (strategy (3))	D-statistic & Calibration slope (strategy (1))	D-statistic & Calibration slope (strategy (2))	D-statistic & Calibration slope (strategy (3))
1	0.842	0.334	0.325	0.349	0.677	0.663	0.704
2	0.827	0.303	0.307	0.324	0.662	0.664	0.668
3	0.702	0.218	0.237	0.240	0.661	0.688	0.691
4	0.762	0.300	0.294	0.297	0.782	0.777	0.779
5	0.834	0.469	0.468	0.468	0.668	0.661	0.660
6	0.807	0.438	0.451	0.442	0.750	0.764	0.754
7	0.612	0.276	0.276	0.274	0.817	0.821	0.832
8	0.812	0.199	0.166	0.183	0.625	0.573	0.599

Strategy (1): Baseline hazard estimated in external validation dataset.
Strategy (2): Average baseline hazard from developed model.
Strategy (3): Baseline hazard from country included in the development, closest in proximity.

Supplementary material 5: Trivariate random-effects meta-analysis results of calibration and discrimination performance for the breast cancer model excluding Study 3, for implementation strategy (2)

Strategy	Validation statistic	Pooled estimate (95% CI)	95% prediction interval	I-squared	Estimate of τ	Joint probability of good* performance in a new population
Strategy (2) including Study 3: Develop using Royston-Parmar model and apply with the estimated average baseline hazard from developed model	Calibration slope	0.994 (0.835 to 1.153)	0.411 to 1.577	98%	0.224	0.22
	C-statistic	0.711 (0.691 to 0.732)	0.662 to 0.76	43%	0.017	
	D-statistic	0.332 (0.212 to 0.452)	-0.08 to 0.744	88%	0.157	
Strategy (2) excluding Study 3: Develop using Royston-Parmar model and apply with the estimated average baseline hazard from developed model	Calibration slope	0.999 (0.883 to 1.114)	0.594 to 1.404	95%	0.146	0.32
	C-statistic	0.712 (0.688 to 0.735)	0.650 to 0.773	52%	0.021	
	D-statistic	0.372 (0.256 to 0.490)	-0.014 to 0.760	85%	0.138	

* defined by a C-statistic≥0.7 and a calibration slope between 0.9 and 1.1