A guide to systematic review and meta-analysis of prediction model performance

Thomas P A Debray,[1,2] Johanna A A G Damen,[1,2] Kym I E Snell,[3] Joie Ensor,[3] Lotty Hooft,[1,2] Johannes B Reitsma,[1,2] Richard D Riley,[3] Karel G M Moons[1,2]

[1]Cochrane Netherlands, University Medical Center Utrecht, PO Box 85500 Str 6.131, 3508 GA Utrecht, Netherlands

[2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500 Str 6.131, 3508 GA Utrecht, Netherlands

[3]Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, UK

Correspondence to: T P A Debray T.Debray@umcutrecht.nl

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2017;356:i6460
http://dx.doi.org/10.1136/bmj.i6460

Accepted: 25 November 2016

Validation of prediction models is highly recommended and increasingly common in the literature. A systematic review of validation studies is therefore helpful, with meta-analysis needed to summarise the predictive performance of the model being validated across different settings and populations. This article provides guidance for researchers systematically reviewing and meta-analysing the existing evidence on a specific prediction model, discusses good practice when quantitatively summarising the predictive performance of the model across studies, and provides recommendations for interpreting meta-analysis estimates of model performance. We present key steps of the meta-analysis and illustrate each step in an example review, by summarising the discrimination and calibration performance of the EuroSCORE for predicting operative mortality in patients undergoing coronary artery bypass grafting.

Systematic reviews and meta-analysis are an important—if not the most important—source of information for evidence based medicine.[1] Traditionally, they aim to summarise the results of publications or reports of primary treatment studies and (more recently) of primary diagnostic test accuracy studies. Compared to therapeutic intervention and diagnostic test accuracy studies, there is limited guidance on the conduct of systematic reviews and meta-analysis of primary prognosis studies.

A common aim of primary prognostic studies concerns the development of so-called prognostic prediction models or indices. These models estimate the individualised probability or risk that a certain condition will occur in the future by combining information from multiple prognostic factors from an individual. Unfortunately, there is often conflicting evidence about the predictive performance of developed prognostic prediction models. For this reason, there is a growing demand for evidence synthesis of (external validation) studies assessing a model's performance in new individuals.[2] A similar issue relates to diagnostic prediction models, where the validation performance of a model for predicting the risk of a disease being already present is of interest across multiple studies.

Previous guidance papers regarding methods for systematic reviews of predictive modelling studies have addressed the searching,[3-5] design,[2] data extraction, and critical appraisal[6 7] of primary studies. In this paper, we provide further guidance for systematic review and for meta-analysis of such models. Systematically reviewing the predictive performance of one or more prediction models is crucial to examine a model's predictive ability across different study populations, settings, or locations,[8-11] and to evaluate the need for further adjustments or improvements of a model.

Although systematic reviews of prediction modelling studies are increasingly common,[12-17] researchers often refrain from undertaking a quantitative synthesis or meta-analysis of the predictive performance of a specific model. Potential reasons for this pitfall are concerns about the quality of included studies, unavailability of relevant summary statistics due to incomplete reporting,[18] or simply a lack of methodological guidance.

Based on previous publications, we therefore first describe how to define the systematic review question, to identify the relevant prediction modelling studies from the literature[3 5] and to critically appraise the identified studies.[6 7] Additionally, and not yet addressed in previous publications, we provide guidance on which predictive performance measures could be extracted from the primary studies, why they are important, and how to deal with situations when they are missing or poorly reported. The need to extract aggregate results and information from published studies provides unique challenges that are not faced when individual participant data are available, as described recently in *The BMJ*.[19]

SUMMARY POINTS

Systematic review of the validation studies of a prediction model might help to identify whether its predictions are sufficiently accurate across different settings and populations

Efforts should be made to restore missing information from validation studies and to harmonise the extracted performance statistics

Heterogeneity should be expected when summarising estimates of a model's predictive performance

Meta-analysis should primarily be used to investigate variation across validation study results

We subsequently discuss how to quantitatively summarise the extracted predictive performance estimates and investigate sources of between-study heterogeneity. The different steps are summarised in figure 1, some of which are explained further in different appendices. We illustrate each step of the review using an empirical example study—that is, the synthesis of studies validating predictive performance of the additive European system for cardiac operative risk evaluation (EuroSCORE). Here onwards, we focus on systematic review and meta-analysis of a specific prognostic prediction model. All guidance can, however, similarly be applied to the meta-analysis of diagnostic prediction models. We focus on statistical criteria of good performance (eg, in terms of discrimination and calibration) and highlight other clinically important measures of performance (such as net benefit) in the discussion.

Empirical example

As mentioned earlier, we illustrate our guidance using a published review of studies validating EuroSCORE.[13] This prognostic model aims to predict 30 day mortality in patients undergoing any type of cardiac surgery (appendix 1). It was developed by a European steering group in 1999 using logistic regression in a dataset from 13 302 adult patients undergoing cardiac surgery under cardiopulmonary bypass. The previous review identified 67 articles assessing the performance of the EuroSCORE in patients that were not used for the development of the model (external validation studies).[13] It is important to evaluate whether the predictive performance of EuroSCORE is adequate, because poor performance could eventually lead to poor decision making and thereby affect patient health.

In this paper, we focus on the validation studies that examined the predictive performance of the so-called additive EuroSCORE system in patients undergoing (only) coronary artery bypass grafting (CABG). We included a total of 22 validations, including more than 100 000 patients from 20 external validation studies and from the original development study (appendix 2).

Steps of the systematic review

Formulating the review question and protocol

As for any other type of biomedical research, it is strongly recommended to start with a study protocol describing the rationale, objectives, design, methodology, and statistical considerations of the systematic review.[20] Guidance for formulating a review question for systematic review of prediction models has recently been provided by the CHARMS checklist (checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies).[6] This checklist addresses a modification (PICOTS) of the PICO system (population, intervention, comparison, and outcome) used in therapeutic studies, and additionally considers timing (that is, at which time point and over what time period the outcome is predicted) and setting (that is, the role or setting of the prognostic model). More information on the different items is provided in box 1 and appendix 3.

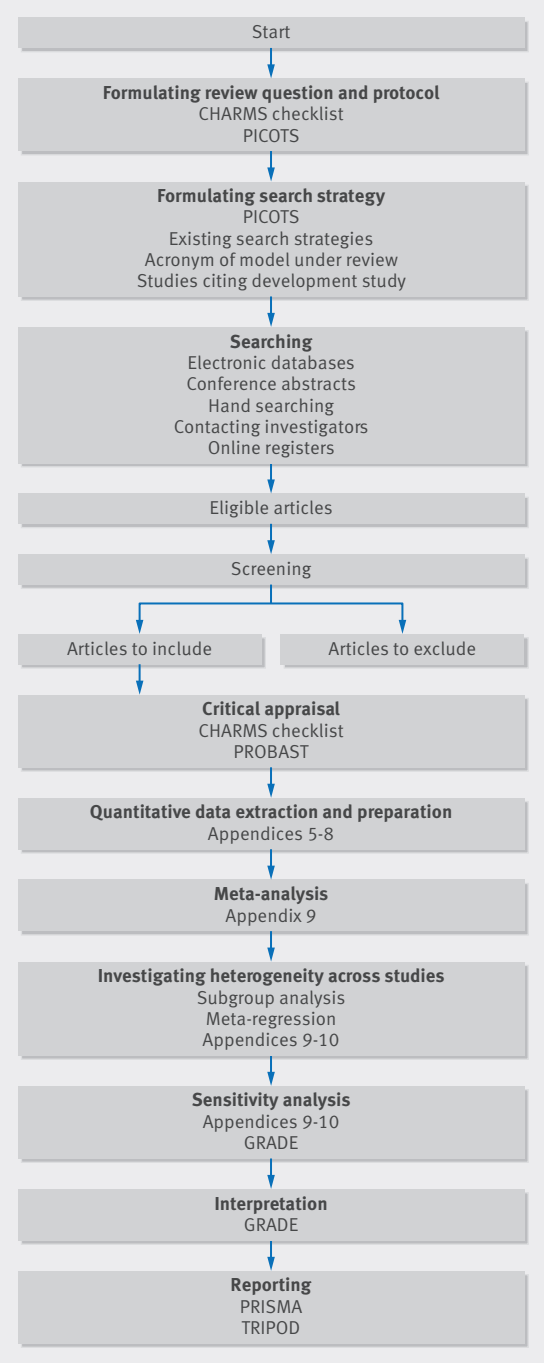


Fig 1 | Flowchart for systematically reviewing and, if considered appropriate, meta-analysis of the validation studies of a prediction model. CHARMS=checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies; PROBAST=prediction model risk of bias assessment tool; PICOTS=population, intervention, comparator, outcome(s), timing, setting; GRADE=grades of recommendation, assessment, development, and evaluation; PRISMA=preferred reporting items for systematic reviews and meta-analyses; TRIPOD=transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

Case study

The formal review question was as follows: to what extent is the additive EuroSCORE able to predict all

Box 1: PICOTS system

The PICOTS system, as presented in the CHARMS checklist,[6] describes key items for framing the review aim, search strategy, and study inclusion and exclusion criteria. The items are explained below in brief, and applied to our case study:

- Population—define the target population in which the prediction model will be used. In our case study, the population of interest comprises patients undergoing coronary artery bypass grafting.
- Intervention (model)—define the prediction model(s) under review. In the case study, the focus is on the prognostic additive EuroSCORE model.
- Comparator—if applicable, one can address competing models for the prognostic model under review. The existence of alternative models was not considered in our case study.
- Outcome(s)—define the outcome(s) of interest for which the model is validated. In our case study, the outcome was defined as all cause mortality. Papers validating the EuroSCORE model to predict other outcomes such as cardiovascular mortality were excluded.
- Timing—specifically for prognostic models, it is important to define when and over what time period the outcome is predicted. Here, we focus on all cause mortality at 30 days, predicted using preoperative conditions.
- Setting—define the intended role or setting of the prognostic model. In the case study, the intended use of the EuroSCORE model was to perform risk stratification in the assessment of cardiac surgical results, such that operative mortality could be used as a valid measure of quality of care.

cause mortality at 30 days in patients undergoing CABG? The question is primarily interested in the predictive performance of the original EuroSCORE, and not how it performs after it has been recalibrated or adjusted in new data.

Formulating the search strategy

When reviewing studies that evaluate the predictive performance of a specific prognostic model, it is important to ensure that the search strategy identifies all publications that validated the model for the target population, setting, or outcomes at interest. To this end, the search strategy should be formulated according to aforementioned PICOTS of interest. Often, the yield of search strategies can further be improved by making use of existing filters for identifying prediction modelling studies[3-5] or by adding the name or acronym of the model under review. Finally, it might help to inspect studies that cite the original publication in which the model was developed.[15]

Case study

We used a generic search strategy including the terms "EuroSCORE" and "Euro SCORE" in the title and abstract. The search resulted in 686 articles. Finally, we performed a cross reference check in the retrieved articles, and identified one additional validation study of the additive EuroSCORE.

Critical appraisal

The quality of any meta-analysis of a systematic review strongly depends on the relevance and methodological quality of included studies. For this reason, it is important to evaluate their congruence with the review question, and to assess flaws in the design, conduct, and analysis of each validation study. This practice is also

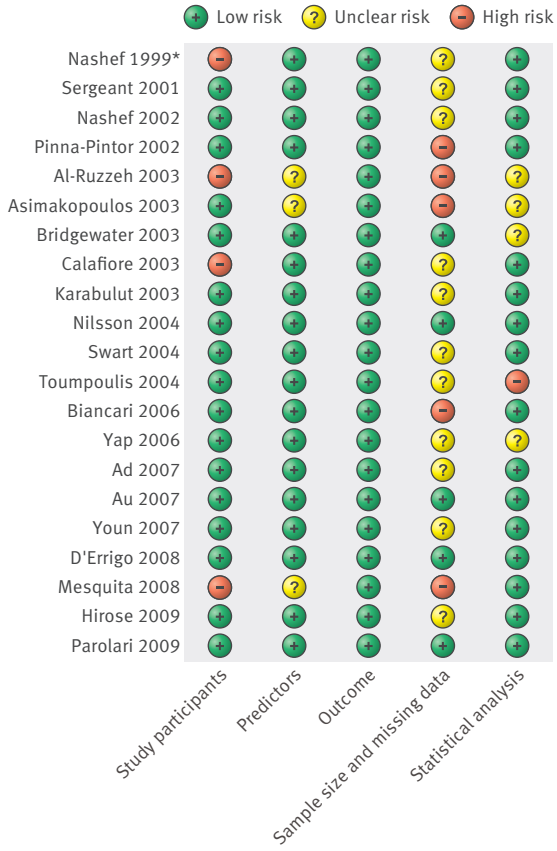


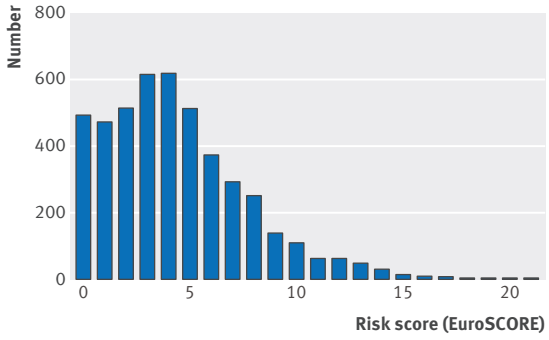
Fig 2 | Overall judgment for risk of bias of included articles in the case study (predictive performance of the EuroSCORE for all cause mortality at 30 days in patients undergoing coronary artery bypass grafting). Study references listed in appendix 2. Study participants domain=design of the included validation study, and inclusion and exclusion of its participants; predictors domain=definition, timing, and measurement of predictors in the validation study (it also assesses whether predictors have not been measured and were therefore omitted from the model in the validation study); outcome domain=definition, timing, and measurement of predicted outcomes; sample size and missing data domain=number of participants in the validation study and exclusions owing to missing data; statistical analysis domain=validation methods (eg, whether the model was recalibrated before validation). Note that there are two validations presented in Nashef 2002; the same scores apply to both model validations. *Original development study (split sample validation)

recommended by Cochrane, and can be implemented using the CHARMS checklist,[6] and, in the near future, using the prediction model risk of bias assessment tool (PROBAST).[7]

Case study

Using the CHARMS checklist and a preliminary version of the PROBAST tool, we critically appraised the risk of bias of each retrieved validation study of the EuroSCORE, as well as of the model development study. Most (n=14) of the 22 validation studies were of low or unclear risk of bias (fig 2). Unfortunately, several validation studies did not report how missing data were

Example 1
We consider here the situation where the distribution of the linear predictor is provided in a figure. In the figure below we can approximate the number of patients for each value of the additive EuroSCORE: 0 (n≈470), 1 (n≈450), 2 (n≈500), 3 (n≈600), 4 (n≈600), 5 (n≈500), 6 (n≈380), 7 (n≈300), 8 (n≈250), 9 (n≈170), 10 (n≈100), 11 (n≈50), 12 (n≈50), 13 (n≈40), 14 (n≈20), 15 (n≈10), and n=1 for the remaining scores. The standard deviation (SD) can then directly be calculated from the corresponding list of 4511 values, and corresponds to 3.



Example 2
Sometimes, the distribution of the linear predictor is reported separately for different subgroups. For instance, in one paper the mean (μ) and standard deviation of the additive EuroSCORE was reported for 3440 patients undergoing on-pump coronary bypass grafting (3.26±2.45) and for 1140 patients undergoing off-pump coronary artery bypass grafting (3.94±2.57). The mean and standard deviation for the linear predictor of the combined group is then given as[25]:

$$\mu = \frac{3440 \times 3.26 + 1140 \times 3.94}{(3440 + 1140)} = 3.43$$

$$SD = \sqrt{\frac{(3440-1) \times 2.45^2 + (1140-1) \times 2.57^2 + \frac{3440 \times 1140}{3440 + 1140}(3.26^2 + 3.94^2 - 2 \times 3.26 \times 3.94)}{3440 + 1140 - 1}}$$

$$= 2.50$$

Example 3
Another validation study reported the median EuroSCORE as 8 (interquartile range 6-11). If we assume that the additive EuroSCORE is normally distributed, the width of the interquartile range is approximately given as 1.35 standard deviations. Hence, we have:

$$SD = \frac{11 - 6}{1.35} = 3.70$$

Fig 3 | Estimation of the standard deviation of the linear predictor as a way to quantify case mix variation within a study

handled (n=13) or performed complete case analysis (n=5). We planned a sensitivity analysis that excluded all validation studies with high risk of bias for at least one domain (n=8).[21]

Quantitative data extraction and preparation

To allow for quantitative synthesis of the predictive performance of the prediction model under study, the necessary results or performance measures and their precision need to be extracted from each model validation study report. The CHARMS checklist can be used for this guidance. We briefly highlight the two most common statistical measures of predictive performance, discrimination and calibration, and discuss how to deal with unreported or inconsistent reporting of these performance measures.

Discrimination

Discrimination refers to a prediction model's ability to distinguish between patients developing and not developing the outcome, and is often quantified by the concordance (C) statistic. The C statistic ranges from 0.5 (no discriminative ability) to 1 (perfect discriminative ability). Concordance is most familiar from logistic regression models, where it is also known as the area under the receiver operating characteristics (ROC) curve. Although C statistics are the most common reported estimates of prediction model performance, they can still be estimated from other reported quantities when missing. Formulas for doing this are presented in appendix 7 (along with their standard errors), and implement the transformations that are needed for conducting the meta-analysis (see meta-analysis section below).

The C statistic of a prediction model can vary substantially across different validation studies. A common cause for heterogeneity in reported C statistics relates to differences between studied populations or study designs.[8 22] In particular, it has been demonstrated that the distribution of patient characteristics (so-called case mix variation) could substantially affect the discrimination of the prediction model, even when the effects of all predictors (that is, regression coefficients) remain correct in the validation study.[22] The more similarity that exists between participants of a validation study (that is, a more homogeneous or narrower case mix), the less discrimination can be achieved by the prediction model.

Therefore, it is important to extract information on the case mix variation between patients for each included validation study,[8] such as the standard deviation of the key characteristics of patients, or of the linear predictor (fig 3). The linear predictor is the weighted sum of the values of the predictors in the validation study, where the weights are the regression coefficients of the prediction model under investigation.[23] Heterogeneity in reported C statistics might also appear when predictor effects differ across studies (eg, due to different measurement methods of predictors), or when different definitions (or different derivations) of the C statistic have been used. Recently, several concordance measures have been proposed that allow to disentangle between different sources of heterogeneity.[22 24] Unfortunately, these measures are currently rarely reported.

Case study

We found that the C statistic of the EuroSCORE was reported in 20 validations (table 1). When measures of uncertainty were not reported, we approximated the standard error of the C statistic (seven studies) using the equations provided in appendix 7 (fig 4). Furthermore, for each validation, we extracted the standard deviation of the age distribution and of the linear predictor of the additive EuroSCORE to help quantify the case mix variation in each study. When such information could not be retrieved, we estimated the standard deviation from reported ranges or histograms (fig 3).[26]

Table 1 | Details of the 22 validations of the additive EuroSCORE to predict overall mortality at 30 days

Study (country, enrolment year)	Validation study results							
	Total sample size	Observed deaths (total No)	Expected deaths as predicted by the model (total No)	C statistic*	EuroSCORE†		Calibration plot	Calibration table‡
					Mean	SD		
Nashef 1999 (8 countries, 1995)§	1497	70.6	72.4	0.7590	—	—	Absent	Present
Sergeant 2001 (Belgium, 1997-2000)	2051	81	101.8	0.83 (0.03)	5	4	Present	Present
Nashef 2002 (USA, 1995)	153 397	—	—	0.78	—	—	Absent	Absent
Nashef 2002 (USA, 1998-99)	—	—	—	0.75	—	—	Absent	Absent
Pinna-Pintor 2002 (Italy, 1993-94)	418	7	—	0.806	2.32	2.0	Present	Absent
Al-Ruzzeh 2003 (UK, 1996-2000)¶	1907	26	49.6	0.77 (0.67 to 0.86)	—	—	Absent	Present
Asimakopoulos 2003 (UK, 1993-99)¶	4654	152	137	0.76 (0.72 to 0.80)	—	—	Present	Absent
Bridgewater 2003 (UK, 1999-2002)	8572	144	257	0.75	3.0	2.48**	Present	Absent
Calafiore 2003 (Italy, 1994-2001)	1020	46	76.4	—	7.8	—	Absent	Present
Karabulut 2003 (Turkey, 1999-2001)	912	10	29.5	0.828	3.23	2.62††	Absent	Present
Nilsson 2004 (1996-2001)	4497	85	85	0.84 (0.80 to 0.88)	4.28**	3.11**	Present	Present
Swart 2004 (South Africa)	574	21	22.39	0.80	—	—	Absent	Absent
Toumpoulis 2004 (USA, 1992-2002)	3760	103	—	0.75 (0.70 to 0.79)	5.38	2.99	Absent	Present
Biancari 2006 (Finland, 1992-93)	917	5	—	0.856 (0.706 to 1.006)	2.22**	2.09**	Absent	Present
Yap 2006 (Australia, 2001-05)	5592	112	237.66	0.82	4.25	3.43††	Absent	Present
Ad 2007 (USA, 2001-04)	3125	57	134.38	—	4.3	3.2	Absent	Absent
Au 2007 (Hong Kong, 1999-2005)	1247	36	49.88	0.76 (0.68 to 0.85)	4.0	3.3	Absent	Absent
Youn 2007 (Korea, 2002-06)	757	10	34.2	0.72 (0.57 to 0.87)	4.5	2.8	Absent	Present
D'Errigo 2008 (Italy, 2002-04)	30 610	777	—	0.773 (0.755 to 0.791)	—	—	Present	Absent
Mesquita 2008 (Brazil, 2005-07)	144	7	7.34	0.702 (0.485 to 0.919)	4	3	Absent	Absent
Hirose 2009 (Japan, 1991-2006)	1522	14	—	0.890	2.9	2.2	Present	Present
Parolari 2009 (Italy, 1999-2007)	3440	29	108.88	0.808 (0.723 to 0.892)	3.26	2.45	Absent	Absent

SD=standard deviation.
*Data are standard error or 95% confidence intervals.
†Scores for risk factors in the EuroSCORE are added to give an approximate percentage predicted mortality, such that expected deaths≈total sample size×mean EuroSCORE/100 and mean EuroSCORE≈expected deaths×100/total sample size.
‡Presented with total number of observed deaths and total number of expected deaths as predicted by the model across different risk stratums.
§Original development study. Results are based on split sample validation. No external validation was applied.
¶The effect of pulmonary hypertension was not incorporated into the calculation of the additive EuroSCORE because the corresponding predictor was not measured.
**Estimated from a histogram or calibration table (fig 3).
††Standard deviation was estimated from a 95% confidence interval (appendix 7).

Calibration

Calibration refers to a model's accuracy of predicted risk probabilities, and indicates the extent to which expected outcomes (predicted from the model) and observed outcomes agree. It is preferably reported graphically with expected outcome probabilities plotted against observed outcome frequencies (so-called calibration plots, see appendix 4), often across tenths of predicted risk.[23] Also for calibration, reported performance estimates might vary across different validation studies. Common causes for this are differences in overall prognosis (outcome incidence). These differences might appear because of differences in healthcare quality and delivery, for example, with screening programmes in some countries identifying disease at an earlier stage, and thus apparently improving prognosis in early years compared to other countries. This again emphasises the need to identify studies and participants relevant to the target population, so that a meta-analysis of calibration performance is relevant.

Summarising estimates of calibration performance is challenging because calibration plots are most often not presented, and because studies tend to report different types of summary statistics in calibration.[12 27] Therefore, we propose to extract information on the total number of observed (O) and expected (E) events, which are statistics most likely to be reported or derivable (appendix 7). The total O:E ratio provides a rough indication of the overall model calibration (across the entire range of predicted risks). The total O:E ratio is strongly related to the calibration in the large (appendix 5), but that is rarely reported. The O:E ratio might also be available in subgroups, for example, defined by tenths of predicted risk or by particular groups of interest (eg, ethnic groups, or regions). These O:E ratios could also be extracted, although it is unlikely that all studies will report the same subgroups. Finally, it would be helpful to also extract and summarise estimates of the calibration slope.

Case study

Calibration of the additive EuroSCORE was visually assessed in seven validation studies. Although the total O:E ratio was typically not reported, it could be calculated from other information for 19 of the 22 included validations. For nine of these validation studies, it was also possible to extract the proportion of observed outcomes across different risk strata of the additive EuroSCORE (appendix 8). Measures of uncertainty were often not reported (table 1). We therefore approximated the standard error of the total O:E ratio (19 validation studies) using the equations provided in appendix 7. The forest plot displaying the study specific results is presented in figure 4. The calibration slope was not

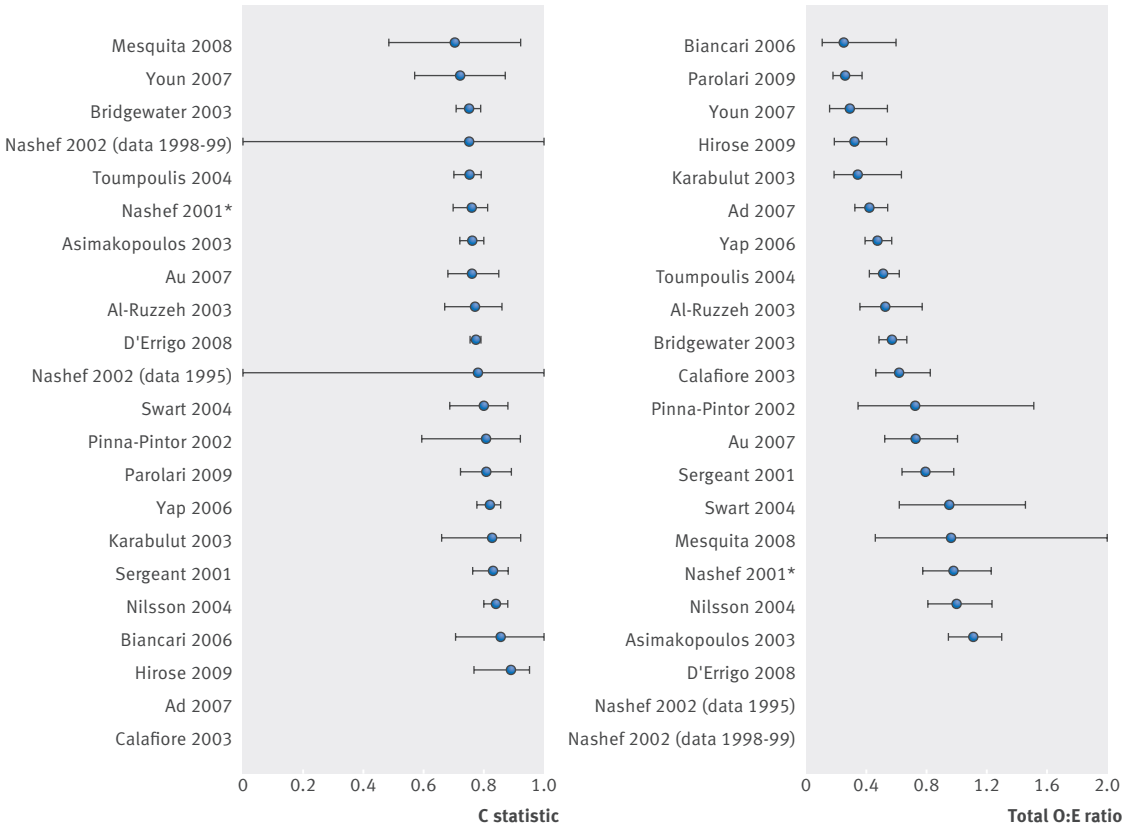


Fig 4 | **Forest plots of extracted performance statistics of the additive EuroSCORE in the case study (to predict all cause mortality at 30 days in patients undergoing coronary artery bypass grafting). Part A shows forest plot of study specific C statistics (all 95% confidence intervals estimated on the logit scale); part B shows forest plot of study specific total O:E ratios (where O=total number of observed deaths and E=total number of expected deaths as predicted by the model; when missing, 95% confidence intervals were approximated on the log scale using the equations from appendix 7). *Performance in the original development study (split sample validation)**

reported for any validation study and could not be derived using other information.

Performance of survival models

Although we focus on discrimination and calibration measures of prediction models with a binary outcome, similar performance measures exist for prediction models with a survival (time to event) outcome. Caution is, however, warranted when extracting reported C statistics because different adaptations have been proposed for use with time to event outcomes.[9 28 29] We therefore recommend to carefully evaluate the type of reported C statistic and to consider additional measures of model discrimination.

For instance, the D statistic gives the log hazard ratio of a model's predicted risks dichotomised at the median value, and can be estimated from Harrell's C statistic when missing.[30] Finally, when summarising the calibration performance of survival models, it is recommended to extract or calculate O:E ratios for particular (same) time points because they are likely to differ across time. When some events remain unobserved, owing to censoring, the total number of events and the observed outcome risk at particular time points should be derived (or approximated) using Kaplan-Meier estimates or Kaplan-Meier curves.

Meta-analysis

Once all relevant studies have been identified and corresponding results have been extracted, the retrieved estimates of model discrimination and calibration can be summarised into a weighted average. Because validation studies typically differ in design, execution, and thus case-mix, variation between their results are unlikely to occur by chance only.[8 22] For this reason, the meta-analysis should usually allow for (rather than ignore) the presence of heterogeneity and aim to produce a summary result (with its 95% confidence interval) that quantifies the average performance across studies. This can be achieved by implementing a random (rather than a fixed) effects meta-analysis model (appendix 9). The meta-analysis then also yields an estimate of the between-study standard deviation, which directly quantifies the extent of heterogeneity across studies.[19] Other meta-analysis models have also been proposed, such as by Pennells and colleagues, who suggest weighting by the number of events in each study because this is the principal determinant of study precision.[31] However, we recommend to use traditional random effects models where the weights are based on the within-study error variance. Although it is common to summarise estimates of model discrimination and calibration separately, they can also jointly be

synthesised using multivariate meta-analysis.[9] This might help to increase precision of summary estimates, and to avoid exclusion of studies for which relevant estimates are missing (eg, discrimination is reported but not calibration).

To further interpret the relevance of any between-study heterogeneity, it is also helpful to calculate an approximate 95% prediction interval (appendix 9). This interval provides a range for the potential model performance in a new validation study, although it will usually be very wide if there are fewer than 10 studies.[32] It is also possible to estimate the probability of good performance when the model is applied in practice.[9] This probability can, for instance, indicate the likelihood of achieving a certain C statistic in a new population. In case of multivariate meta-analysis, it is even possible to define multiple criteria of good performance. Unfortunately, when performance estimates substantially vary across studies, summary estimates might not be very informative. Of course, it is also desirable to understand the cause of between-study heterogeneity in model performance, and we return to this issue in the next section.

Some caution is warranted when summarising estimates of model discrimination and calibration. Previous studies have demonstrated that extracted C statistics[33-35] and total O:E ratios[33] should be rescaled before meta-analysis to improve the validity of its underlying assumptions. Suggestions for the necessary transformations are provided in appendix 7. Furthermore, in line with previous recommendations, we propose to adopt restricted maximum likelihood (REML) estimation and to use the Hartung-Knapp-Sidik-Jonkman (HKSJ) method when calculating 95% confidence intervals for the average performance, to better account for the uncertainty in the estimated between-study heterogeneity.[36 37] The HKSJ method is implemented in several meta-analysis software packages, including the metareg module in Stata (StataCorp) and the metafor package in R (R Foundation for Statistical Computing).

Case study

To summarise the performance of the EuroSCORE, we performed random effects meta- analyses with REML estimation and HKSJ confidence interval derivation. For model discrimination, we found a summary C statistic of 0.79 (95% confidence interval 0.77 to 0.81; approximate 95% prediction interval 0.72 to 0.84). The probability of so-called good discrimination (defined as a C statistic >0.75) was 89%. For model calibration, we found a summary O:E ratio of 0.53. This implies that, on average, the additive EuroSCORE substantially overestimates the risk of all cause mortality at 30 days. The weighted average of the total O:E ratio is, however, not very informative because 95% prediction intervals are rather wide (0.19 to 1.46). This problem is also illustrated by the estimated probability of so-called good calibration (defined as an O:E ratio between 0.8 and 1.2), which was only 15%. When jointly meta-analysing discrimination and calibration performance, we found similar summary estimates for the C statistic and total O:E ratio. The joint probability of good performance (defined as C statistic >0.75 and O:E ratio between 0.8 and 1.2), however, decreased to 13% owing to the large extent of miscalibration. Therefore, it is important to investigate potential sources of heterogeneity in the calibration performance of the additive EuroSCORE model.

Investigating heterogeneity across studies

When the discrimination or calibration performance of a prediction model is heterogeneous across validation studies, it is important to investigate potential sources of heterogeneity. This may help to understand under what circumstances the model performance remains adequate, and when the model might require further improvements. As mentioned earlier, the discrimination and calibration of a prediction model can be affected by differences in the design[38] and in populations across the validation studies, for example, owing to changes in case mix variation or baseline risk.[8 22]

In general, sources of heterogeneity can be explored by performing a meta-regression analysis where the dependent variable is the (transformed) estimate of the model performance measure.[39] Study level or summarised patient level characteristics (eg, mean age) are then used as explanatory or independent variables. Alternatively, it is possible to summarise model performance across different clinically relevant subgroups. This approach is also known as subgroup analysis and is most sensible when there are clearly definable subgroups. This is often only practical if individual participant data are available.[19]

Key issues that could be considered as modifiers of model performance are differences in the heterogeneity between patients across the included validation studies (difference case mix variation),[8] differences in study characteristics (eg, in terms of design, follow-up time, or outcome definition), and differences in the statistical analysis or characteristics related to selective reporting and publication (eg, risk of bias, study size). The regression coefficient obtained from a meta-regression analysis describes how the dependent variable (here, the logit C statistic or log O:E ratio) changes between subgroups of studies in case of a categorical explanatory variable or with one unit increase in a continuous explanatory variable. The statistical significance measure of the regression coefficient is a test of whether there is a (linear) relation between the model's performance and the explanatory variable. However, unless the number of studies is reasonably large (>10), the power to detect a genuine association with these tests will usually be low. In addition, it is well known that meta-regression and subgroup analysis are prone to ecological bias when investigating summarised patient level covariates as modifiers of model performance.[40]

Case study

To investigate whether population differences generated heterogeneity across the included validation

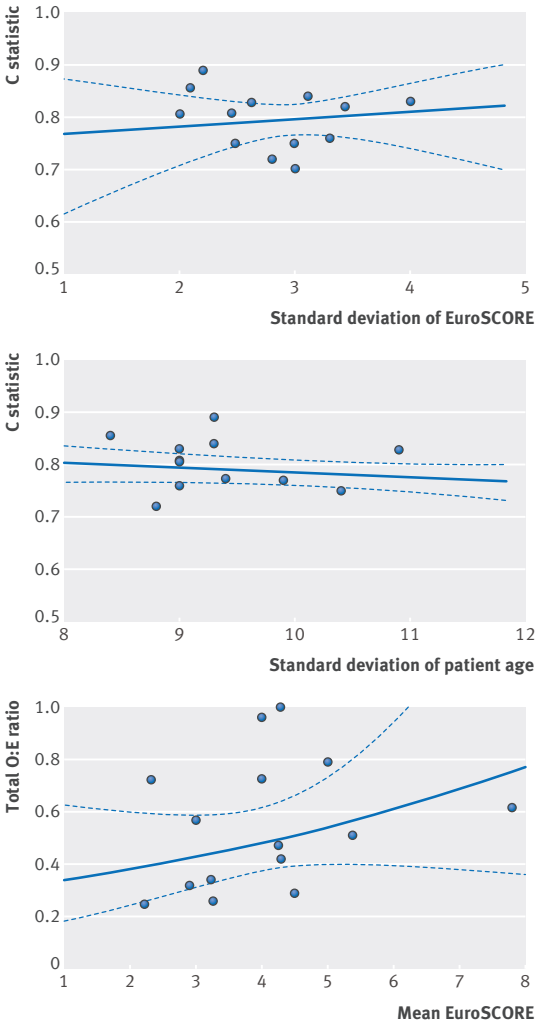


Fig 5 | Results from random effects meta-regression models in the case study (predictive performance of the EuroSCORE for all cause mortality at 30 days in patients undergoing coronary artery bypass grafting). Solid lines=regression lines; dashed lines=95% confidence intervals; dots=included validation studies

studies, we performed several meta-regression analyses (fig 5 and appendix 10). We first evaluated whether the summary C statistic was related to the case mix variation, as quantified by the spread of the EuroSCORE in each validation study, or related to the spread of patient age. We then evaluated whether the summarised O:E ratio was related to the mean EuroSCORE values, year of study recruitment, or continent. Although the power was limited to detect any association, results suggest that the EuroSCORE tends to overestimate the risk of early mortality in low risk populations (with a mean EuroSCORE value <6). Similar results were found when we investigated the total O:E ratio across different subgroups, using the reported calibration tables and histograms within the included validation studies (appendix 8). Although year of study recruitment and continent did not significantly influence the calibration, we found that miscalibration was more problematic in (developed) countries with low mortality rates (appendix 10). The C statistic did not appear to differ importantly as the standard deviation of the EUROSCORE or age distribution increased.

Overall, we can conclude that the additive EuroSCORE fairly discriminates between mortality and survival in patients undergoing CABG. Its overall calibration, however, is quite poor because predicted risks appear too high in low risk patients, and the extent of miscalibration substantially varies across populations. Not enough information is available to draw conclusions on the performance of EuroSCORE in high risk patients. Although it has been suggested that overprediction likely occurs due to improvements in cardiac surgery, we could not confirm this effect in the present analyses.

Sensitivity analysis

As for any meta-analysis, it is important to show that results are not distorted by low quality validation studies. For this reason, key analyses should be repeated for the studies at lower and higher risk of bias.

Case study

We performed a subgroup analysis by excluding those studies at high risk of bias, to ascertain their effect (fig 2). Results in table 2 indicate that this approach yielded similar summary estimates of discrimination and calibration as those in the full analysis of all studies.

Reporting and presentation

As for any other type of systematic review and meta-analysis, it is important to report the conducted research in sufficient detail. The PRISMA statement (preferred reporting items for systematic reviews and meta-analyses)[41] highlights the key issues for reporting of meta-analysis of intervention studies, which are also generally relevant for meta-analysis of model validation studies. If meta-analysis of individual participant data (IPD) has been used, then PRISMA-IPD will also be helpful.[42] Furthermore, the TRIPOD statement (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis)[23 43] provides several recommendations for the reporting of studies developing, validating, or updating a prediction model, and can be considered here as well. Finally, use of the GRADE approach (grades of recommendation, assessment, development, and evaluation) might help to interpret the results of the systematic review and to present the evidence.[21]

As illustrated in this article, researchers should clearly describe the review question, search strategy, tools used for critical appraisal and risk of bias assessment, quality of the included studies, methods used for data extraction and meta-analysis, data used for meta-analysis, and corresponding results and their uncertainty. Furthermore, we recommend to report details on the relevant study populations (eg, using the mean and standard deviation of the linear predictor) and to present summary estimates with confidence intervals and, if appropriate, prediction intervals.

Table 2 | Results from the case study (predictive performance of the EuroSCORE for all cause mortality at 30 days in patients undergoing coronary artery bypass grafting) after excluding studies with high risk of bias

Meta-analysis	Performance	Risk of bias	No of included studies	Summary estimate	95% confidence interval	95% prediction interval
Univariate*	C statistic	Low/unclear/high	18	0.78	0.76 to 0.80	0.73 to 0.83
Univariate*	O:E ratio	Low/unclear/high	19	0.55	0.43 to 0.69	0.20 to 1.53
Bivariate*	C statistic	Low/unclear/high	20	0.79	0.77 to 0.80	0.73 to 0.83
Bivariate*	O:E ratio	Low/unclear/high	20	0.55	0.44 to 0.68	0.20 to 1.47
Univariate	C statistic	Low/unclear/high	17	0.79	0.77 to 0.81	0.72 to 0.84
Univariate	O:E ratio	Low/unclear/high	18	0.53	0.42 to 0.67	0.19 to 1.46
Bivariate	C statistic	Low/unclear/high	19	0.79	0.77 to 0.81	0.73 to 0.84
Bivariate	O:E ratio	Low/unclear/high	19	0.53	0.42 to 0.66	0.20 to 1.40
Univariate	C statistic	Low/unclear	13	0.80	0.77 to 0.82	0.73 to 0.85
Univariate	O:E ratio	Low/unclear	13	0.49	0.36 to 0.67	0.16 to 1.50
Bivariate	C statistic	Low/unclear	14	0.80	0.77 to 0.82	0.73 to 0.85
Bivariate	O:E ratio	Low/unclear	14	0.48	0.37 to 0.64	0.17 to 1.40
Univariate	C statistic	Low	4	0.80	0.73 to 0.85	0.66 to 0.89
Univariate	O:E ratio	Low	3	0.57	0.10 to 3.33	0.02 to 19.15
Bivariate	C statistic	Low	4	0.80	0.74 to 0.84	0.70 to 0.87
Bivariate	O:E ratio	Low	4	0.52	0.19 to 1.40	0.06 to 4.09

Results are based on random effects meta-analyses with REML estimation and HKSJ confidence interval derivation. For bivariate meta-analyses, we assumed zero within-study correlation between the reported C statistic and the total O:E ratio.
*Includes results from the split sample validation of the development study of the additive EuroSCORE.

Finally, it might be helpful to report probabilities of good performance separately for each performance measure, because researchers can then decide which criteria are most relevant for their situation.

Concluding remarks

In this article, we provide guidance on how to systematically review and quantitatively synthesize the predictive performance of a prediction model. Although we focused on systematic review and meta-analysis of a prognostic model, all guidance can similarly be applied to the meta-analysis of a diagnostic prediction model. We discussed how to define the systematic review question, identify the relevant prediction model studies from the literature, critically appraise the identified studies, extract relevant summary statistics, quantitatively summarise the extracted estimates, and investigate sources of between-study heterogeneity.

Meta-analysis of a prediction model's predictive performance bears many similarities to other types of meta-analysis. However, in contrast to synthesis of randomised trials, heterogeneity is much more likely in meta-analysis of studies assessing the predictive performance of a prediction model, owing to the increased variation of eligible study designs, increased inclusion of studies with different populations, and increased complexity of required statistical methods. When substantial heterogeneity occurs, summary estimates of model performance can be of limited value. For this reason, it is paramount to identify relevant studies through a systematic review, assess the presence of important subgroups, and evaluate the performance the model is likely to yield in new studies.

Although several concerns can be resolved by aforementioned strategies, it is possible that substantial between-study heterogeneity remains and can only be addressed by harmonising and analysing the study individual participant data.[19] Previous studies have demonstrated that access to individual participant data might also help to retrieve unreported performance measures (eg, calibration slope), estimate the within-study correlation between performance measures,[9] avoid continuity corrections and data transformations, further interpret model generalisability,[8 19 22 31] and tailor the model to populations at hand.[44]

Often, multiple models exist for predicting the same condition in similar populations. In such situations, it could be desirable to investigate their relative performance. Although this strategy has already been adopted by several authors, caution is warranted in the absence of individual participant data. In particular, the lack of head-to-head comparisons between competing models and the increased likelihood of heterogeneity across validation studies renders comparative analyses highly prone to bias. Further, it is well known that performance measures such as the C statistic are relatively insensitive to improvements in predictive performance. We therefore believe that summary performance estimates might often be of limited value, and that a meta-analysis should rather focus on assessing their variability across relevant settings and populations. Formal comparisons between competing models are possible (eg, by adopting network meta-analysis methods) but appear most useful for exploratory purposes.

Finally, the following limitations need to be considered in order to fully appreciate this guidance. Firstly, our empirical example demonstrates that the level of reporting in validation studies is often poor. Although the quality of reporting has been steadily improving over the past few years, it will often be necessary to restore missing information from other quantities. This strategy might not always be reliable, such that sensitivity analyses remain paramount in any meta-analysis. Secondly, the statistical methods we discussed in this article are most applicable when meta-analysing

the performance results from prediction models developed with logistic regression. Although the same principles apply to survival models, the level of reporting tends to be even less consistent because many more statistical choices and multiple time points need to be considered. Thirdly, we focused on frequentist methods for summarising model performance and calculating corresponding prediction intervals. Bayesian methods have, however, been recommended when predicting the likely performance in a future validation study.[45] Lastly, we mainly focused on statistical measures of model performance, and did not discuss how to meta-analyse clinical measures of performance such as net benefit.[46] Because these performance measures are not frequently reported and typically require subjective thresholds, summarising them appears difficult without access to individual participant data. Nevertheless, further research on how to meta-analyse net benefit estimates would be welcome.

In summary, systematic review and meta-analysis of prediction model performance could help to interpret the potential applicability and generalisability of a prediction model. When the meta-analysis shows promising results, it may be worthwhile to obtain individual participant data to investigate in more detail how the model performs across different populations and subgroups.[19 44]

We thank *The BMJ* editors and reviewers for their helpful feedback on this manuscript.

Contributors: KGMM, TPAD, JBR, and RDR conceived the paper objectives. TPAD prepared a first draft of this article, which was subsequently reviewed in multiple rounds by JAAGD, JE, KIES, LH, RDR, JBR, and KGMM. TPAD and JAAGD undertook the data extraction and statistical analyses. TPAD, JAAGD, RDR, and KGMM contributed equally to the paper. All authors approved the final version of the submitted manuscript. TPAD is guarantor. All authors had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Funding: Financial support received from the Cochrane Methods Innovation Funds Round 2 (MTH001F) and the Netherlands Organization for Scientific Research (91617050 and 91810615). This work was also supported by the UK Medical Research Council Network of Hubs for Trials Methodology Research (MR/L004933/1- R20). RDR was supported by an MRC partnership grant for the PROGnosis RESearch Strategy (PROGRESS) group (grant G0902393). None of the funding sources had a role in the design, conduct, analyses, or reporting of the study or in the decision to submit the manuscript for publication.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: support from the Cochrane Methods Innovation Funds Round 2, Netherlands Organization for Scientific Research, and the UK Medical Research Council for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

1 Khan K, Kunz R, Kleijnen J, et al. Systematic reviews to support evidence-based medicine: how to review and apply findings of healthcare research. CRC Press, 2nd ed, 2011.
2 Steyerberg EW, Moons KGM, van der Windt DA, et al. PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10:e1001381. doi:10.1371/journal.pmed.1001381.
3 Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews [correction in: *PLoS One* 2012;7(7): doi/10.1371/annotation/96bdb520-d704-45f0-a143-43a48552952e]. *PLoS One* 2012;7:e32844. doi:10.1371/journal.pone.0032844.
4 Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R. Hedges Team. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. *AMIA Annu Symp Proc* 2003:728-32.
5 Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc* 2001;8:391-7. doi:10.1136/jamia.2001.0080391.
6 Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11:e1001744. doi:10.1371/journal.pmed.1001744.
7 Wolff R, Whiting P, Mallett S, et al. PROBAST: a risk of bias tool for prediction modelling studies. Cochrane Colloquium Vienna. 2015.
8 Debray TPA, 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. doi:10.1016/j.jclinepi.2014.06.018.
9 Snell KI, Hua H, Debray TP, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2016;69:40-50. doi:10.1016/j.jclinepi.2015.05.009.
10 Altman DG, Royston P. What do we mean by validating a prognostic model?*Stat Med* 2000;19:453-73. doi:10.1002/(SICI)1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5.
11 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130:515-24. doi:10.7326/0003-4819-130-6-199903160-00016.
12 Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013;66:268-77. doi:10.1016/j.jclinepi.2012.06.020.
13 Siregar S, Groenwold RHH, de Heer F, Bots ML, van der Graaf Y, van Herwerden LA. Performance of the original EuroSCORE. *Eur J Cardiothorac Surg* 2012;41:746-54. doi:10.1093/ejcts/ezr285.
14 Echouffo-Tcheugui JB, Batty GD, Kivimäki M, Kengne AP. Risk models to predict hypertension: a systematic review. *PLoS One* 2013;8:e67370. doi:10.1371/journal.pone.0067370.
15 Tzoulaki I, Liberopoulos G, Ioannidis JPA. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* 2009;302:2345-52. doi:10.1001/jama.2009.1757.
16 Eichler K, Puhan MA, Steurer J, Bachmann LM. Prediction of first coronary events with the Framingham score: a systematic review. *Am Heart J* 2007;153:722-31, 731.e1-8. doi:10.1016/j.ahj.2007.02.027.
17 Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak* 2006;6:38. doi:10.1186/1472-6947-6-38.
18 Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;14:40. doi:10.1186/1471-2288-14-40.
19 Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;353:i3140. doi:10.1136/bmj.i3140.
20 Peat G, Riley RD, Croft P, et al. PROGRESS Group. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. *PLoS Med* 2014;11:e1001671. doi:10.1371/journal.pmed.1001671.
21 Iorio A, Spencer FA, Falavigna M, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. *BMJ* 2015;350:h870. doi:10.1136/bmj.h870.
22 Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010;172:971-80. doi:10.1093/aje/kwq223.
23 Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1-73. doi:10.7326/M14-0698.
24 van Klaveren D, Gönen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings. *Stat Med* 2016;35:4136-52. doi:10.1002/sim.6997.
25 Higgins JPT, Green S. Combining Groups. http://handbook.cochrane.org/chapter_7/7_7_3_8_combining_groups.htm, 2011.
26 Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. *BMC Med Res Methodol* 2005;5:13. doi:10.1186/1471-2288-5-13.
27 Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9:1-12. doi:10.1371/journal.pmed.1001221.
28 Austin PC, Pencinca MJ, Steyerberg EW. Predictive accuracy of novel risk factors and markers: A simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model. *Stat Methods Med Res* 2015;0962280214567141.

29 Blanche P, Dartigues JF, Jacqmin-Gadda H. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biom J* 2013;55:687-704. doi:10.1002/bimj.201200045.
30 Jinks RC, Royston P, Parmar MKB. Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. *BMC Med Res Methodol* 2015;15:82. doi:10.1186/s12874-015-0078-y.
31 Pennells L, Kaptoge S, White IR, Thompson SG, Wood AM. Emerging Risk Factors Collaboration. Assessing risk prediction models using individual participant data from multiple studies. *Am J Epidemiol* 2014;179:621-32. doi:10.1093/aje/kwt298.
32 Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011;342:d549. doi:10.1136/bmj.d549.
33 Snell KIE. Development and application of statistical methods for prognosis research. PhD thesis, School of Health and Population Sciences, Birmingham, UK, 2015.
34 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. doi:10.1186/1471-2288-14-5.
35 Gengsheng Qin , Hotilovac L. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Stat Methods Med Res* 2008;17:207-21. doi:10.1177/0962280207087173.
36 IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* 2014;14:25. doi:10.1186/1471-2288-14-25.
37 Cornell JE, Mulrow CD, Localio R, et al. Random-effects meta-analysis of inconsistent effects: a time for change. *Ann Intern Med* 2014;160:267-70. doi:10.7326/M13-2886.
38 Ban JW, Emparanza JI, Urreta I, Burls A. Design Characteristics Influence Performance of Clinical Prediction Rules in Validation: A Meta-Epidemiological Study. *PLoS One* 2016;11:e0145779. doi:10.1371/journal.pone.0145779.
39 Deeks JJ, Higgins JPT, Altman DG. Chapter 9. *Analysing data and undertaking meta-analyses*.Cochrane Collaboration, 2011.
40 Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Anti-Lymphocyte Antibody Induction Therapy Study Group. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Stat Med* 2002;21:371-87. doi:10.1002/sim.1023.
41 Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;339:b2700. doi:10.1136/bmj.b2700.
42 Stewart LA, Clarke M, Rovers M, et al. PRISMA-IPD Development Group. Preferred reporting items for a systematic review and meta- analysis of individual participant data: The PRISMA-IPD Statement. *JAMA* 2015;313:1657-65. doi:10.1001/jama.2015.3656.
43 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162:55-63. doi:10.7326/M14-0697.
44 Debray TPA, Riley RD, Rovers MM, Reitsma JB, Moons KG. Cochrane IPD Meta-analysis Methods group. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. *PLoS Med* 2015;12:e1001886. doi:10.1371/journal.pmed.1001886.
45 Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res* 2001;10:277-303. doi:10.1191/096228001678227794.
46 Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016;352:i6. doi:10.1136/bmj.i6

Appendix: Supplementary material