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

Thomas P.A. Debray†, Johanna A. A. G. Damen†, Kym I. E. Snell, Joie Ensor, Lotty Hooft, Johannes B. Reitsma, Richard D. Riley†, Karel G. M. Moons†

Corresponding author:
Thomas PA Debray
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht
P.O. Box 85500
Str. 6.131
3508 GA Utrecht
The Netherlands
T.Debray@umcutrecht.nl
+31 88 75 680 26

Thomas P A Debray
Assistant professor
Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, 3508 GA Utrecht, The Netherlands

Johanna A A G Damen
PhD fellow
Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, 3508 GA Utrecht, The Netherlands

Kym I E Snell
Research Fellow
Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, ST5 5BG, The United Kingdom

Joie Ensor
Research Fellow
Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, ST5 5BG, The United Kingdom

Lotty Hooft
Associate professor
Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, 3508 GA Utrecht, The Netherlands

Johannes B Reitsma
Associate professor
Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, 3508 GA Utrecht, The Netherlands

Richard D Riley
Professor
Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, ST5 5BG, The United Kingdom

Karel G M Moons
Professor
Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, 3508 GA Utrecht, The Netherlands

Keywords: meta-analysis, systematic review, prediction model, external validation, prognostic model

† Equal contribution

Approximate word count: 4850

Abstract

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. The aims of this article are (1) to provide guidance for systematically reviewing and meta-analysing the existing evidence on a specific prediction model, (2) to discuss 'good practice' when quantitatively summarizing the predictive performance of the model across studies and (3) to provide recommendations for interpreting meta-analysis estimates of model performance. We present key steps of the meta-analysis and illustrate each step in an exemplar review where we summarize the discrimination and calibration performance of the EuroSCORE for predicting operative mortality in patients undergoing coronary artery bypass grafting.

Summary points

- Systematically reviewing the validation studies of a prediction model may 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 harmonize the extracted performance statistics.
- Heterogeneity should be expected when summarizing estimates of a model's predictive performance.
- Meta-analysis should primarily be used to investigate variation across validation study results.

Introduction

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 summarize 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 very limited guidance for 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 subjects [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 modeling studies 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 modeling 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 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 (IPD) are available, as described recently in the BMJ [19]. We subsequently discuss how to quantitatively summarize the extracted predictive performance estimates and investigate sources of between-study heterogeneity. The different steps are summarized in Figure 1. We illustrate each step of the review using an empirical example study, i.e., 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 (e.g. in terms of discrimination and calibration) and highlight other clinically important measures of performance (such as net-benefit) in the discussion.

Empirical example

We illustrate our guidance using a published review of studies validating the additive European system for cardiac operative risk evaluation (EuroSCORE) [13]. This prognostic model aims to predict 30-day mortality in patients undergoing any type of cardiac surgery (details in Appendix 1). It was developed by a European steering group in 1999 using logistic regression in a dataset from 13 302 adult patients

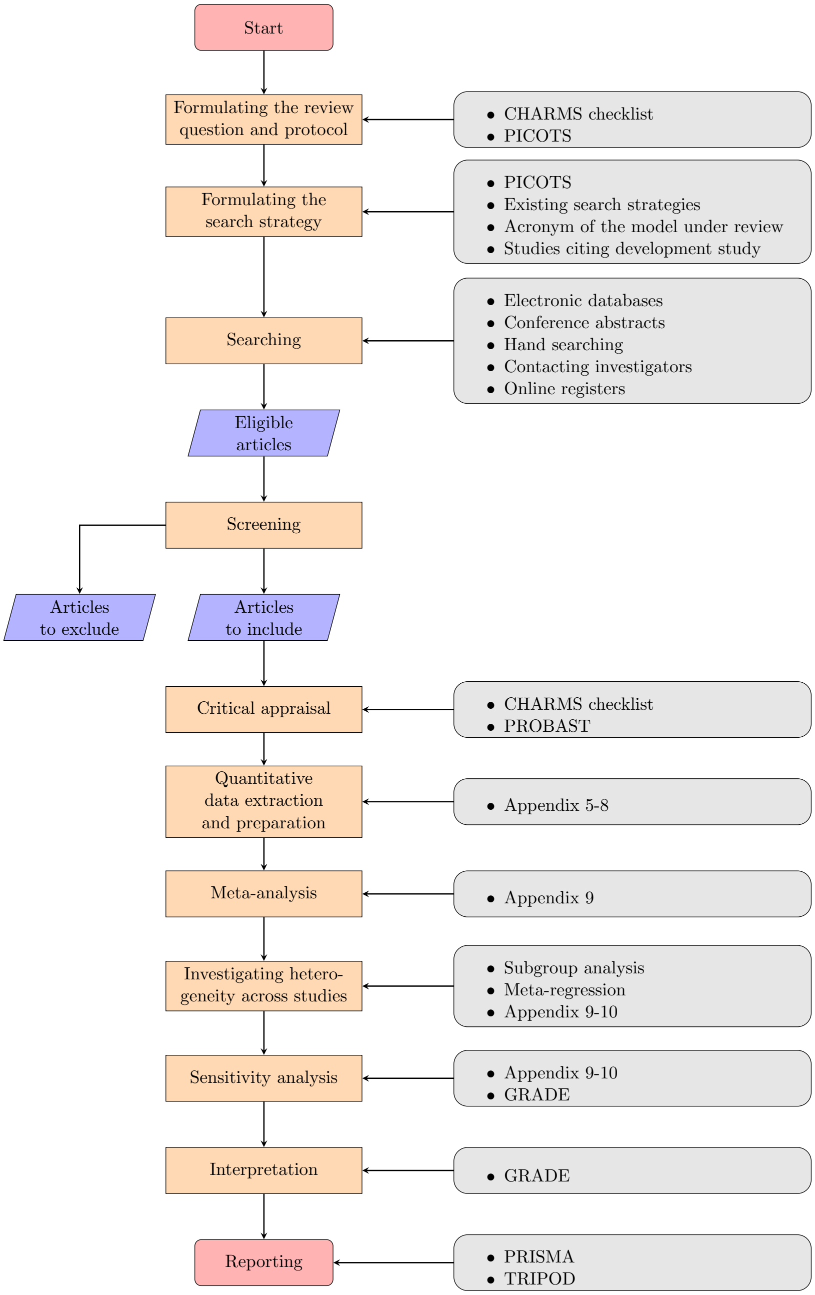


Fig 1. Flow chart for systematically reviewing and, if considered appropriate, meta-analysis of the validation studies of a prediction model
CHARMS = CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies; PROBAST = Prediction model risk of bias assessment 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

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, as poor performance may 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 one hundred thousand patients from 20 external validation studies and from the original development study (Appendix 2).

Patient involvement

No patients were involved in setting the research question or the outcome measures, nor were they involved in developing plans for design or implementation of the study. No patients were asked to advise on interpretation or writing up of results. There are no plans to disseminate the results of the research to study participants or the relevant patient community.

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 Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) [6]. This checklist addresses a modification (PICOTS) of the PICO (population, intervention, comparison, outcome) system used in therapeutic studies, and additionally considers timing (i.e. at which time point and over what time period the outcome is predicted) and setting (i.e. the role or setting of the prognostic model). More information on the different items is provided in Box 1 and in Appendix 3.

Case study: The formal review question was as follows: to what extent is the additive EuroSCORE able to predict 30-day all-cause mortality in patients undergoing CABG? We are 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.

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. 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 CABG.
- **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 may 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. We here focus on 30-day all cause mortality, 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.

Box 1. The PICOTS system.

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 may 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 upon 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 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: Based on 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

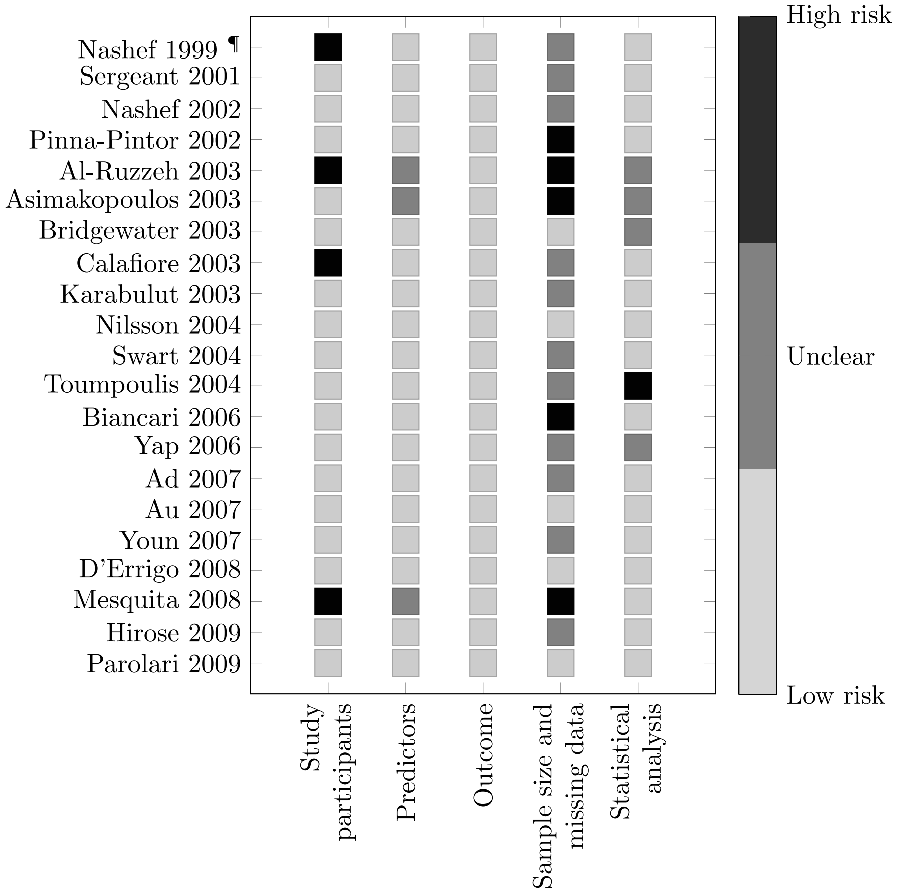


Fig 2. Overall judgement for risk of bias of the included articles

The domain “study participants” relates to the design of the included validation study and the in- and exclusion of its participants. The domain “predictors” relates to the definition, timing and measurement of the predictors in the validation study. It also assesses if predictors have not been measured and were therefore omitted from the model in the validation study. The domain “outcome” relates to the definition, timing and measurement of predicted outcomes. The domain “sample size and missing data” relates to the amount of subjects in the validation study and exclusions due to missing data. Finally, the domain “statistical analysis” relates to the validation methods, e.g. whether the model was recalibrated before validation. Note that there are 2 validations presented in Nashef 2002. The same scores apply to both model validations.

¶ Original development study (split sample validation) risk of bias (Figure 2). Unfortunately, several validation studies did not report how missing data were handled ($n = 13$) or performed complete case analysis ($n = 5$). We planned a sensitivity analysis where all validation studies with high risk of bias for at least one domain ($n = 8$) were excluded [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.

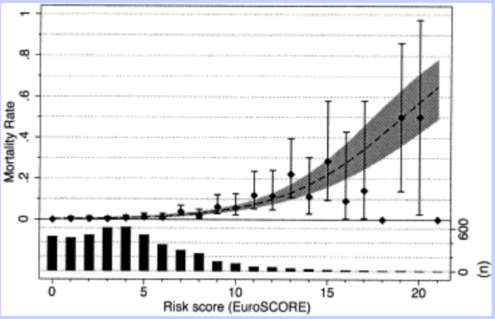
Discrimination

Discrimination refers to a prediction model's ability to distinguish between subjects 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 section below).

It is important to realize that the c-statistic of a prediction model may substantially vary 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 subject characteristics (so-called case-mix variation) may substantially influence the discrimination of the prediction model, even when the effects of all predictors (i.e. regression coefficients) remain 'correct' in the validation study [22]. The more similarity exists between the subjects of a validation study (i.e. more homogeneous or narrower case-mix), the less discrimination can be achieved by the prediction model. Hence, 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 subject characteristics and/or of the linear predictor (Box 2). 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 may also appear when predictor effects differ across studies (e.g. 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.

Example 1

We here consider 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 ($n \approx 470$), 1 ($n \approx 450$), 2 ($n \approx 500$), 3 ($n \approx 600$), 4 ($n \approx 600$), 5 ($n \approx 500$), 6 ($n \approx 380$), 7 ($n \approx 300$), 8 ($n \approx 250$), 9 ($n \approx 170$), 10 ($n \approx 100$), 11 ($n \approx 50$), 12 ($n \approx 50$), 13 ($n \approx 40$), 14 ($n \approx 20$), 15 ($n \approx 10$) and $n = 1$ for the remaining scores. The SD can then directly be calculated from the corresponding list of 4 511 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 SD 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 SD for the linear predictor of the combined group is then given as [25]:

$$\begin{aligned}\mu &= \frac{3440 \times 3.26 + 1140 \times 3.94}{(3440 + 1140)} = 3.43 \\ \text{SD} &= \sqrt{\frac{(3440-1) \times 2.45^2 + (1140-1) \times 2.57^2 + \frac{3440 \times 1140}{3440+1140}(3.26^2 + 3.94^2 - 2 \times 3.26 \times 3.94)}{3340 + 1140 - 1}} \\ &= 2.50\end{aligned}$$

Example 3

Another validation study reported the median EuroSCORE (8) with an interquartile range (6 to 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:

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

Box 2. Estimating the standard deviation (SD) of the linear predictor as a way to quantify case-mix variation within a study

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 ($n = 7$ studies) using the equations provided in Appendix 7 (Figure 3). Furthermore, for each validation, we extracted the standard deviation (SD) 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 [26] (see Box 2).

Calibration

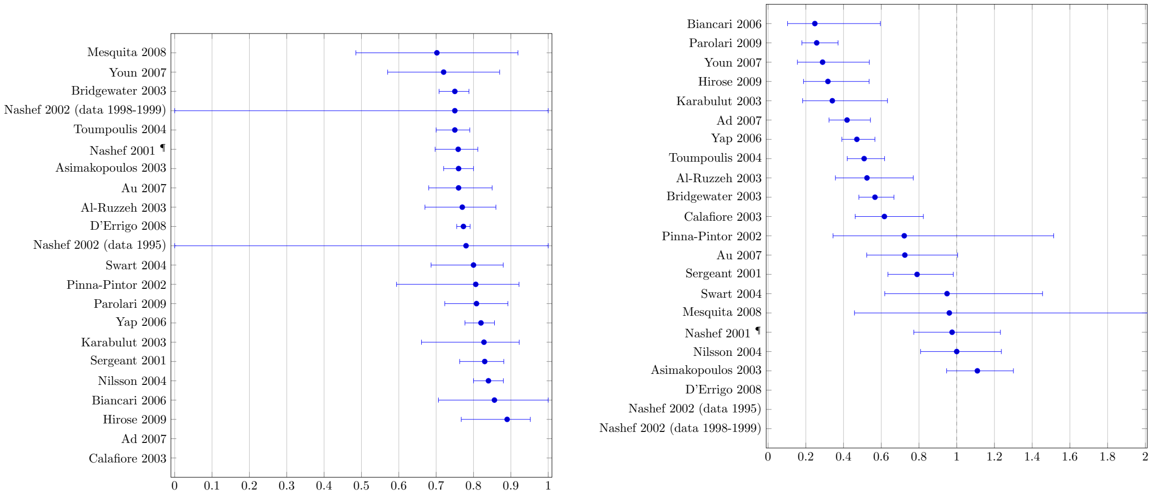
Calibration refers to a model's accuracy of predicted risk probabilities, and indicates the extent to which expected (predicted from the model) and observed outcomes agree. It is preferably reported graphically with expected outcome probabilities plotted against observed outcome frequencies (so-called calibration plots, see Appendix 4), often across tenths of predicted risk [23]. Also for calibration, reported performance estimates may vary across different validation studies. Common causes for this are differences in overall prognosis (outcome incidence). These may for instance appear due to differences in health care 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.

Unfortunately, summarizing 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]. We therefore 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. Sometimes, the O:E may also be available in subgroups, for example defined by tenths of predicted risk or by particular groups of interest (e.g. 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 summarize estimates of the calibration slope.

Case study: Calibration of the additive EuroSCORE was visually assessed in 7 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 9 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 ($n = 19$ validation studies) using the equations provided in Appendix 7. The resulting forest plot displaying the study-specific results is depicted in Figure 3. The calibration slope was not 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 dichotomized at the median value, and can be estimated from Harrell's c-statistic when missing [30]. Finally, when summarizing the calibration performance of survival models, it is recommended to extract or calculate O:E ratios for particular (same)



(i) Forest plot of the study-specific c-statistics. All 95% confidence intervals were estimated on the logit scale.

(ii) Forest plot of the study-specific total O:E ratios. When missing, 95% confidence intervals were approximated on the log scale using the equations from Appendix 7.

Fig 3. Forest plot of the extracted performance statistics of the additive EuroSCORE. ¶ Performance in the original development study (split sample validation).

time-points as they are likely to differ across time. When some events remain unobserved due to censoring, the total number of events and the observed outcome risk at particular time-points should be derived (or approximated) using 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 summarized 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 Pennells *et al.*, who suggest weighting by the number of events in each study as 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 summarize estimates of model discrimination and calibration separately, they can also jointly be synthesized using multivariate meta-analysis [9]. This may help to increase precision of summary estimates, and to avoid exclusion of studies for which relevant estimates are missing (e.g. 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, though 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 may 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 summarizing 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 prior to 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, College Station, Texas) and the `metafor` package in R (R Foundation for Statistical Computing, Vienna, Austria).

Case study: To summarize 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 with a 95% confidence interval ranging from 0.77 to 0.81 and an approximate 95% prediction interval ranging from 0.72 to 0.84. The probability of “good” discrimination (defined as a c-statistic above 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 over-estimates the risk of all-cause 30-day mortality. 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 “good” calibration (defined as $0.8 \leq \mathrm{O} : \mathrm{E} \leq 1.2$), 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 a c-statistic above 0.75 together with $0.8 \leq$ O:E ≤ 1.2), however, decreased to 13% due to the large extent of mis-calibration. For this reason, 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 may require further improvements. As said, 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 due to changes in case-mix variation and/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 summarized patient-level characteristics (e.g. mean age) are then used as explanatory or independent variables. Alternatively, it is possible to summarize 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 IPD are available [19].

Key issues that could be considered as modifiers of model performance are differences in the heterogeneity between subjects across the included validation studies (difference case-mix variation) [8], differences in study characteristics (e.g. in terms of design, follow-up time or in outcome definition) and differences in the statistical analysis or characteristics related to selective reporting and publication (e.g. 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) relationship 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 summarized patient-level covariates as modifiers of model performance [40].

Case study: To investigate whether population differences generated heterogeneity across the included validation studies, we performed several meta-regression analyses (Figure 4 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 participant age. Afterwards, we evaluated whether the summarized 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 over-estimate the risk of early mortality in low-risk populations (with a mean EuroSCORE value below 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 mis-calibration was more problematic in (developed) countries with low mortality rates (Appendix 10). The c-statistic did not appear to differ importantly as the SD 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 as predicted risks appear too high in low-risk patients and as the extent of mis-calibration 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 over-prediction likely occurs due to improvements in cardiac surgery, we could not confirm this effect in the present analyses.

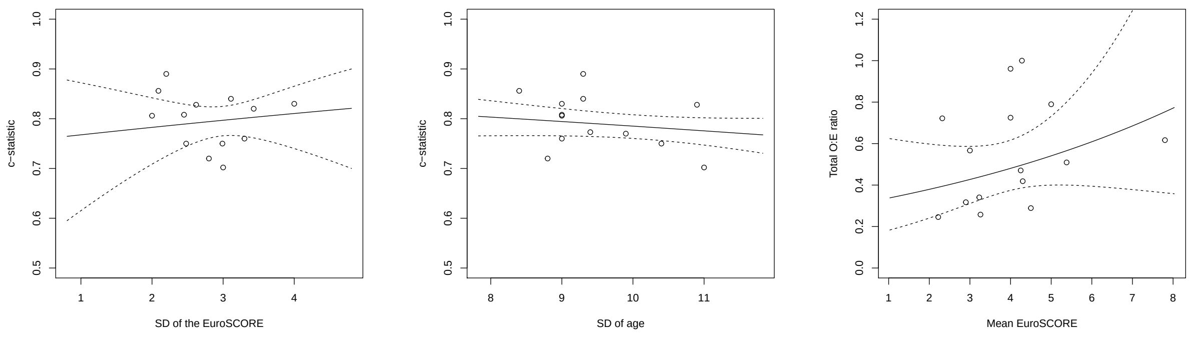


Fig 4. Results from random-effects meta-regression models. Dashed lines indicate the bounds of the 95% confidence interval around the regression line. Dots indicate the included validation studies.

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 impact (see Figure 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 preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement [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 IPD has been used, then PRISMA-IPD will also be helpful [42]. Furthermore, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [23,43] provides several recommendations for the reporting of studies developing, validating, or updating a prediction model, and may be considered here as well. Finally, the Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) may 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, the search strategy, the tools used for critical appraisal and risk of bias assessment, the quality of the included studies, the methods used for data extraction and meta-analysis, the data used for meta-analysis and the corresponding results and their uncertainty. Furthermore, we recommend to report details on the relevant study populations (e.g. using the mean and SD of the linear predictor) and to present summary estimates with confidence intervals and, if appropriate, prediction intervals. Finally, it may be helpful to report probabilities of "good" performance separately for each performance measure, as researchers can then decide which criteria are most relevant for their situation.

Concluding remarks

We provided 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, to identify the relevant prediction model studies from the literature, to critically appraise the identified studies, to extract relevant summary statistics, to quantitatively summarize the extracted estimates and to 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 randomized trials, heterogeneity is much more likely in meta-analysis of studies assessing the predictive performance of a prediction model due to the increased variation of eligible study designs, the increased inclusion of studies with different populations, and the 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, to assess the presence of important subgroups, and to 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 harmonizing and analyzing the study IPD [19]. Previous studies have demonstrated that access to IPD may also help to retrieve unreported performance measures (e.g. calibration slope), to estimate the within-study correlation between performance measures [9], to avoid continuity corrections and data transformations, to further interpret model generalizability [8, 19, 22, 31] and to tailor the model to populations at hand [44].

Often, multiple models exist for predicting the same condition in similar populations. In such situations, it may be desirable to investigate their relative performance. Although this strategy has already been adopted by several authors, caution is warranted in the absence of IPD. 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 may 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, e.g. 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. First, 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 may not always be very reliable, such that sensitivity analyses remain paramount in any meta-analysis. Second, 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 much more statistical choices and multiple time-points need to be considered. Third, we focused on frequentist methods for summarizing model performance and calculating corresponding prediction intervals. Bayesian methods have, however, been recommended when predicting the likely performance in a future validation study [45]. Last but not least, 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, summarizing them appears difficult without access to IPD. 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 may help to interpret the potential applicability and generalizability of a prediction model. When the meta-analysis shows promising results, it may be worthwhile to obtain IPD to investigate in more detail how the model performs across different populations and subgroups [19, 44].

Acknowledgments

We gratefully acknowledge the helpful feedback from the BMJ editors and from the reviewers of this manuscript: Dr. Claudia Kimie Suemoto, Dr. Chen Mao, Dr. Lisa Pennells and Dr. Javier Ballesteros.

Author Contributions

KM, TD, JR and RR conceived the paper objectives. TD prepared a first draft of this article, which was subsequently reviewed in multiple rounds by JD, JE, KS, LH, RR, JR and KM. TD and JD undertook the data extraction and statistical analyses. All authors approved the final version of the submitted manuscript. TD is guarantor.

Funding

We gratefully acknowledge the financial support by the Cochrane Methods Innovation Funds Round 2 (MTH001F) and by the Netherlands Organization for Scientific Research (91617050 and 91810615). This work was also supported by the MRC Network of Hubs for Trials Methodology Research (MR/L004933/1-R20). Finally, RR was supported by an MRC Partnership Grant for the PROGnosis RESearch Strategy (PROGRESS) group (grant reference number: 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 (available on request from the corresponding author) and declare: no support from any organisation 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.

Ethical approval

Not required.

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.

Transparency

The lead authors (the manuscript's guarantors) affirm that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Data sharing

No additional data are available.

Licensing

References

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, London, 2 edition, 2011.

2. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med* 2013;10:e1001381.

3. Geersing GJ, Bouwmeester W, Zuithoff P, et al. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. *PLoS One* 2012;7:e32844.

4. Wong SS, Wilczynski NL, Haynes RB, et al. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. *AMIA Annu Symp Proc* 2003;728–732.

5. Ingui BJ and Rogers MA. Searching for clinical prediction rules in MEDLINE. *Journal of the American Medical Informatics Association* 2001;8:391–397.

6. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Clinical Prediction Modelling Studies: The CHARMS Checklist. *PLoS Med* 2014;11:e1001744.

7. Wolff R, Whiting P, Mallett S, et al. PROBAST: a risk of bias tool for prediction modelling studies. In Cochrane Colloquium Vienna. 2015; .

8. Debray TPA, Vergouwe Y, Koffijberg H, et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015;68:279–289.

9. Snell K, Hui H, Debray T, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2015;69:40–50.

10. Altman DG and Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000; 19:453–473.

11. Justice AC, Covinsky KE, and Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130:515–524.

12. Collins GS, Omar O, Shanyinde M, et al. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013;66:268–277.

13. Siregar S, Groenwold RHH, de Heer F, et al. Performance of the original EuroSCORE. *Eur J Cardiothorac Surg* 2012;41:746–754.

14. Echouffo-Tcheugui JB, Batty GD, Kivimki M, et al. Risk models to predict hypertension: a systematic review. *PLoS One* 2013;8:e67370.

15. Tzoulaki I, Liberopoulos G, and Ioannidis JPA. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* 2009;302:2345–2352.

16. Eichler K, Puhan MA, Steurer J, et al. Prediction of first coronary events with the Framingham score: a systematic review. *Am Heart J* 2007;153:722–731.

17. Perel P, Edwards P, Wentz R, et al. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak* 2006;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.

19. Riley R, Ensor J, Snell K, 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.

20. Peat G, Riley RD, Croft P, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. *PLoS Med* 2014;11:e1001671.

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.

22. Vergouwe Y, Moons KGM, and Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010;172:971–980.

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–W73.

24. van Klaveren D, Gnen M, Steyerberg EW, et al. A new concordance measure for risk prediction models in external validation settings. *Stat Med* 2016;35:4136–4152.

25. Higgins JPT and Green S. Combining Groups. `http://handbook.cochrane.org/chapter_7/7_7_3_8_combining_groups.htm`, 2011.

26. Hozo SP, Djulbegovic B, and Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. *BMC Med Res Methodol* 2005;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.

28. Austin PC, Pencinca MJ, and 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;.

29. Blanche P, Dartigues JF, and Jacqmin-Gadda H. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal* 2013;55:687–704.

30. Jinks RC, Royston P, and Parmar MKB. Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. *BMC Med Res Methodol* 2015;15:82.

31. Pennells L, Kaptoge S, White IR, et al. Assessing Risk Prediction Models Using Individual Participant Data From Multiple Studies. *Am J Epidemiol* 2013;179:621–632.

32. Riley RD, Higgins JPT, and Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011; 342:d549.

33. Snell KIE. Development and application of statistical methods for prognosis research. Ph.D. thesis, School of Health and Population Sciences, Birmingham, United Kingdom, 2015.

34. van Klaveren D, Steyerberg EW, Perel P, et al. Assessing discriminative ability of risk models in clustered data. *BMC Med Res Methodol* 2014;14:5.

35. Qin G and 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;17:207–221.

36. IntHout J, Ioannidis JPA, and 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.

37. Cornell JE, Mulrow CD, Localio R, et al. Random-effects meta-analysis of inconsistent effects: a time for change. *Ann Intern Med* 2014;160:267–270.

38. Ban JW, Emparanza JI, Urreta I, et al. Design Characteristics Influence Performance of Clinical Prediction Rules in Validation: A Meta-Epidemiological Study. *PLoS One* 2016;11:e0145779.

39. Deeks JJ, Higgins JPT, and Altman DG. Analysing data and undertaking meta-analyses, chapter 9. The Cochrane Collaboration, 2011;.

40. Berlin JA, Santanna J, Schmid CH, et al. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Stat Med* 2002;21:371–387.

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.

42. Stewart LA, Clarke M, Rovers M, et al. Preferred reporting items for a systematic review and meta-analysis of individual participant data: The PRISMA-IPD Statement. *JAMA* 2015;313:1657–1665.

43. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162:55–63.

44. Debray TPA, Riley RD, Rovers MM, et al. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS Med* 2015;12:e1001886.

45. Sutton AJ and Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res* 2001;10:277–303.

46. Vickers AJ, Van Calster B, and Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016;352:i6.

Table 1. Details of the 22 validations of the additive EuroSCORE to predict 30-day overall mortality

Study	**Country**	**Enrolment (year)**	**Validation study results**							
			N	O	E	c-statistic	E-M	E-SD	CP	CT
Nashef 1999 ¶	8 countries	1995	1 497	70.6	72.4	0.7590	-	-	○	●
Sergeant 2001	Belgium	1997–2000	2 051	81	101.8	0.83 ±0.03	5	4	●	●
Nashef 2002	US	1995	153 397	-	-	0.78	-	-	○	○
Nashef 2002	US	1998–1999	-	-	-	0.75	-	-	○	○
Pinna-Pintor 2002	Italy	1993–1994	418	7	-	0.806	2.32	2.0	●	○
Al-Ruzzeh 2003 *	UK	1996–2000	1 907	26	49.6	0.77 (0.67; 0.86)	-	-	○	●
Asimakopoulos 2003 *	UK	1993–1999	4 654	152	137	0.76 (0.72; 0.80)	-	-	●	○
Bridgewater 2003	UK	1999–2002	8 572	144	257	0.75	3.0	2.48†	●	○
Calafiore 2003	Italy	1994–2001	1 020	46	76.4	-	7.8	-	○	●
Karabulut 2003	Turkey	1999–2001	912	10	29.5	0.828	3.23	2.62‡	○	●
Nilsson 2004	Sweden	1996–2001	4 497	85	85	0.84 (0.80; 0.88)	4.28†	3.11†	●	●
Swart 2004	South Africa	-	574	21	22.39	0.80	-	-	○	○
Toumpoulis 2004	US	1992–2002	3 760	103	-	0.75 (0.70; 0.79)	5.38	2.99	○	●
Biancari 2006	Finland	1992–1993	917	5	-	0.856 (0.706; 1.006)	2.22†	2.09†	○	●
Yap 2006	Australia	2001–2005	5 592	112	237.66	0.82	4.25	3.43‡	○	●
Ad 2007	US	2001–2004	3 125	57	134.38	-	4.3	3.2	○	○
Au 2007	Hong Kong	1999–2005	1 247	36	49.88	0.76 (0.68; 0.85)	4.0	3.3	○	○
Youn 2007	Korea	2002–2006	757	10	34.2	0.72 (0.57; 0.87)	4.5	2.8	○	●
D'Errigo 2008	Italy	2002–2004	30 610	777	-	0.773 (0.755; 0.791)	-	-	●	○
Mesquita 2008	Brazil	2005–2007	144	7	7.34	0.702 (0.485; 0.919)	4	3	○	○
Hirose 2009	Japan	1991–2006	1 522	14	-	0.890	2.9	2.2	●	●
Parolari 2009	Italy	1999–2007	3 440	29	108.88	0.808 (0.723; 0.892)	3.26	2.45	○	○

N = total sample size; O = total number of observed deaths; E = total number of expected deaths as predicted by the model; E-M = mean EuroSCORE; E-SD = standard deviation of the EuroSCORE; CP = calibration plot (● = present; ○ = absent); CT = calibration table presented with O and E across different risk strata (● = present; ○ = absent).

The symbol ± indicates a standard error, whereas 95% confidence intervals are presented between brackets. Note that the scores for the risk factors in the EuroSCORE are added to give an approximate percentage predicted mortality, such that $E \approx N \times \text{CMS}/100$ and $\text{CMS} \approx E \times 100/N$.

¶: 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 (Box 2).

‡: The standard deviation was estimated from a 95% confidence interval (Appendix 7).

Table 2. Results from the case study: predictive performance of the EuroSCORE

Meta-analysis	Performance	Risk of bias	N	S. Est.	95% CI	95% PI
Univariate ¶	c-statistic	Low / Unclear / High	18	0.78	0.76 – 0.80	0.73 – 0.83
Univariate ¶	O:E ratio	Low / Unclear / High	19	0.55	0.43 – 0.69	0.20 – 1.53
Bivariate ¶	c-statistic	Low / Unclear / High	20	0.79	0.77 – 0.80	0.73 – 0.83
Bivariate ¶	O:E ratio	Low / Unclear / High	20	0.55	0.44 – 0.68	0.20 – 1.47
Univariate	c-statistic	Low / Unclear / High	17	0.79	0.77 – 0.81	0.72 – 0.84
Univariate	O:E ratio	Low / Unclear / High	18	0.53	0.42 – 0.67	0.19 – 1.46
Bivariate	c-statistic	Low / Unclear / High	19	0.79	0.77 – 0.81	0.73 – 0.84
Bivariate	O:E ratio	Low / Unclear / High	19	0.53	0.42 – 0.66	0.20 – 1.40
Univariate	c-statistic	Low / Unclear	13	0.80	0.77 – 0.82	0.73 – 0.85
Univariate	O:E ratio	Low / Unclear	13	0.49	0.36 – 0.67	0.16 – 1.50
Bivariate	c-statistic	Low / Unclear	14	0.80	0.77 – 0.82	0.73 – 0.85
Bivariate	O:E ratio	Low / Unclear	14	0.48	0.37 – 0.64	0.17 – 1.40
Univariate	c-statistic	Low	4	0.80	0.73 – 0.85	0.66 – 0.89
Univariate	O:E ratio	Low	3	0.57	0.10 – 3.33	0.02 – 19.15
Bivariate	c-statistic	Low	4	0.80	0.74 – 0.84	0.70 – 0.87
Bivariate	O:E ratio	Low	4	0.52	0.19 – 1.40	0.06 – 4.09

Results are based on random effects meta-analyses with REML estimation and HKSJ confidence interval derivation. N = number of included studies, S. Est. = summary estimate, CI = confidence interval, PI = prediction interval.

¶ Includes the results from the split sample validation of the development study of the additive EuroSCORE. For bivariate meta-analysis, we assumed zero within-study correlation between the reported c-statistic and the total O:E ratio.