A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes

Thomas P. A. Debray¹,², Johanna A. A. G. Damen¹,², Richard D. Riley³, Kym Snell³, Johannes B. Reitsma¹,², Lotty Hooft¹,², Gary S. Collins⁴ and Karel G. M. Moons¹,²

Abstract
It is widely recommended that any developed – diagnostic or prognostic – prediction model is externally validated in terms of its predictive performance measured by calibration and discrimination. When multiple validations have been performed, a systematic review followed by a formal meta-analysis helps to summarize overall performance across multiple settings, and reveals under which circumstances the model performs suboptimal (alternative poorer) and may need adjustment.

We discuss how to undertake meta-analysis of the performance of prediction models with either a binary or a time-to-event outcome. We address how to deal with incomplete availability of study-specific results (performance estimates and their precision), and how to produce summary estimates of the $c$-statistic, the observed:expected ratio and the calibration slope. Furthermore, we discuss the implementation of frequentist and Bayesian meta-analysis methods, and propose novel empirically based prior distributions to improve estimation of between-study heterogeneity in small samples. Finally, we illustrate all methods using two examples: meta-analysis of the predictive performance of EuroSCORE II and of the Framingham Risk Score. All examples and meta-analysis models have been implemented in our newly developed R package “metamisc”.

Keywords
meta-analysis, aggregate data, evidence synthesis, systematic review, prognosis, validation, prediction, discrimination, calibration
Introduction

In medicine, many decisions require to estimate the risk or probability of an existing disease (diagnosis) or of developing a future outcome that yet has to occur (prognosis). Although having experience and intuition often provide excellent advice, it is increasingly common to quantify such diagnostic and prognostic probabilities through the use of prediction models. These models commonly combine information from multiple findings, such as from history taking, physical examination, and additional testing such as blood, imaging, elektrofysiology, and omics tests, to provide absolute outcome probabilities for a certain individual. Prediction models can, for instance, be used to inform patients and their treating physicians, to decide upon the administration of subsequent testing (diagnostic models) or treatments (prognostic models), or to identify participants for clinical trials\textsuperscript{1,2}. Well known examples are the European system for cardiac operative risk evaluation (EuroSCORE) II to predict mortality after cardiac surgery\textsuperscript{3} and the Framingham Risk Score for predicting coronary heart disease\textsuperscript{4}.

Over the past few decades, prediction modeling studies have become abundant in the medical literature. For the same disease, outcome or the target population often numerous, sometimes hundreds, prediction models have been published\textsuperscript{5}. This practice is clearly undesirable for health-care professionals, guideline developers and patients, as it obfuscates which model to use in which context. More efforts are therefore needed to evaluate the performance of existing models in new settings and populations, and to adjust them if necessary\textsuperscript{6}. In contrast to redevelopment, validation and updating of prediction models allows to (more effectively) account for information already captured in previous studies, and thus to make better use of existing evidence and data at hand\textsuperscript{7,8}.

The evaluation (and revision) of prediction models can be achieved by performing so-called external validation studies. In these studies, the original model is applied to new individuals whose data were not used in the model development. Model performance is then assessed by comparing the predicted and observed outcomes across all individuals, and by calculating measures of discrimination and calibration. Discrimination quantifies a model’s ability to distinguish individuals who experience the outcome from

\textsuperscript{1}Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands.
\textsuperscript{2}Cochrane Netherlands, University Medical Center Utrecht, Utrecht, The Netherlands.
\textsuperscript{3}Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, United Kingdom.
\textsuperscript{4}University of Oxford, Oxford, United Kingdom.

Corresponding author:
Thomas P. A. Debray, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands.
Email: T.Debray@umcutrecht.nl
those who remain event free, while calibration refers to the agreement between observed outcome frequencies and predicted probabilities. Unfortunately, the interpretation of validation study results is often difficult, as changes in prediction model performance are likely to occur due to sampling error, differences in predictor effects, and/or differences in patient spectrum\textsuperscript{9;10}. Furthermore, because validation studies are often conducted using small and local data sets\textsuperscript{11}, they rarely provide evidence about a model’s potential generalizability across different settings and populations. For this reason, it may come as no surprise that for many developed prediction models, numerous authors have (re-)assessed the discrimination and calibration performance. Systematic reviews – ideally including a formal meta-analysis – are thus urgently needed to summarize their evidence and to better understand under what circumstances developed models perform adequately or require further adjustments.

Previous guidance for systematic reviews of prediction model studies mainly addressed formulation of the review aim, searching\textsuperscript{12}, critical appraisal (CHARMS)\textsuperscript{13} and data extraction of relevant studies. There is, however, little guidance on how to quantitatively synthesize the results of external validation studies. For this reason, we recently discussed meta-analysis methods to summarize and examine a model’s predictive performance across different settings and (sub)populations\textsuperscript{14}. These methods mainly focused on (diagnostic) prediction models with a binary outcome, and may therefore have limited value when reviewing (prognostic) models with a time-to-event outcome.

For this reason, we provide a comprehensive statistical framework to meta-analyze performance estimates of both diagnostic and prognostic prediction models, involving either binary or time-to-event outcomes. In particular, we discuss how to extract and restore relevant (and possibly missing) estimates of prediction model performance, and corresponding estimates of uncertainty. We also discuss how to obtain summary estimates of discrimination and calibration performance, even when none of the primary studies reported such estimates. Finally, we discuss the role and implementation of Bayesian methods for meta-analysis, and contrast their use with the more traditional frequentist methods. We illustrate all methods by reanalyzing the data from previously published reviews involving a prediction model with a binary\textsuperscript{3} and with a time-to-event\textsuperscript{4} outcome. All methods have been implemented in the R package metamisc, which is available from https://CRAN.R-project.org/package=metamisc\textsuperscript{15}. This package aims to facilitate meta-analysis of prediction model performance by integrating the proposed statistical methods for data extraction and evidence synthesis.

**Motivating examples**

We here focus on prediction models with either a binary or time-to-event outcome, as these are most common in the medical literature\textsuperscript{16}. Binary outcomes are typically used to represent the current (health) status of an individual, or to model events that occur within a relatively short amount of time. Examples
include the presence or absence of a particular disease, or death after surgery. Conversely, when the event times are of primary interest, binary outcomes need to be modeled together with the time after which they occurred. Examples of these so-called time-to-event outcomes include the time until the onset of breast cancer, the time until cardiovascular death, or the time until development of coronary heart disease (CHD).

We now introduce two motivating prognostic model examples – one for a binary and one for a time-to-event outcome – that will be used to illustrate the statistical methods that are required to extract relevant estimates of a model’s predictive performance from published studies and how to quantitatively summarize these. Details on the corresponding systematic reviews can be found in Table 1.

In the first example, we meta-analyze the predictive performance of the European system for cardiac operative risk evaluation (EuroSCORE) II. This logistic regression model was developed using 16,828 adult patients from 43 countries to predict mortality after cardiac surgery. The corresponding prediction model equations are presented in Supporting Information 1.1. In a recent review, 23 validations of EuroSCORE II were identified in which its predictive performance was examined in patients undergoing cardiac surgery. For each study, the number of patients, their mean age, and the proportion of female gender were extracted, as well as information on mortality, the concordance (c-) statistic, and the EuroSCORE II mean value. All data are available from the R package metamisc.

In the second example, we meta-analyze the predictive performance of the Framingham Risk Score developed by Wilson et al. This Cox proportional hazards regression model was developed in 1998 for predicting the 10-year risk of CHD in the general population free from CHD and not on treatment (Supporting Information 1.2). A recent systematic review identified 23 studies examining the performance of Framingham Wilson for predicting fatal or nonfatal CHD in male populations (Supporting Information 1.2). For each validation, estimates of model calibration and discrimination were extracted, as well as details on the study and population characteristics.

Data extraction and estimating unreported results to facilitate meta-analysis

The two most common statistical measures of predictive performance are described by discrimination and calibration. In a meta-analysis without access to individual participant data (IPD), we are reliant on extracting these performance measures from study publications. However, they are often poorly reported. We now discuss how to retrieve the necessary but not explicitly reported predictive performance estimates from the primary prediction model (validation) studies using other reported data.
**Model 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, both for prediction models for binary outcomes as well as for time-to-event outcomes. The \(c\)-statistic is an estimated conditional probability that for any pair of a subject who experienced and a subject who did not experience the outcome, the predicted risk of an event is higher for the former. Although \(c\)-statistics are frequently reported in external validation studies, when missing they can also be calculated from other reported quantities. For instance, for prediction models with a binary outcome, the \(c\)-statistic is a reparameterization of Somer’s D\(^{19}\), and can also be derived from the distribution of the linear predictor\(^{20}\), Cohen’s effect size\(^{20}\), or several correlation indices\(^{21}\). An overview of relevant equations for this derivation of the \(c\)-statistic from other measures was previously presented\(^{14}\).

For prediction models with a time-to-event outcome, concordance is not uniquely defined and several variations of the \(c\)-statistic have been proposed\(^{22;23}\). These variations typically adopt different strategies to account for ties or censoring, and may thus lead to different values for \(c\). Although validation studies do not commonly report the definition and estimation method of presented \(c\)-statistics, Harrell’s version\(^{21}\) appears to be most widespread and recommended\(^{22}\). Sometimes, discrimination is measured using Royston and Sauerbrei’s \(D\) index, usually referred to as the “\(D\) statistic”\(^{24}\). This index quantifies prognostic separation of survival curves and is closely related to the standard deviation of the prognostic index, with Jinks et al. suggesting an equation to convert \(c\) values to \(D\) values based on empirical evidence\(^{25}\):

\[
D \approx 5.50(c - 0.5) + 10.26(c - 0.5)^3
\]  

(1)

It is also possible to convert \(D\) values to \(c\) values by making distributional assumptions with respect to the standard deviation of the prognostic index (Supporting Information 2.2)\(^{26}\):

\[
c = 2 \int_0^\infty \expit \left( \sqrt{\frac{\pi}{4}} Du \right) \phi(u) \, du
\]  

(2)

where \(\phi(u)\) is the standard normal density function.

Regarding the standard error of the \(c\)-statistic, when missing, it can be approximated from a combination of the reported \(c\)-statistic, the total sample size and the total number of events in the validation study. We here consider a method proposed by Newcombe to estimate the standard error of the \(c\)-statistic\(^{27}\):
with \( c \) the reported \( c \)-statistic, \( n \) the number of observed events, \( m \) the total number of non-events and 
\[ m^* = n^* = \frac{1}{2}(m + n) - 1. \]
For prediction models with a time-to-event outcome, the equation above is applicable for estimating the standard error of Harrell’s \( c \)-statistic. As illustrated in Supporting Information 2, this method is fairly accurate for validation studies involving a binary or time-to-event outcome, as long as the total number of observed events is sufficiently large (above 10).

As we will discuss in the next section, it is important to transform estimates of the \( c \)-statistic and its standard error to the logit scale prior to meta-analysis\(^{28}\). The logit \( c \)-statistic is simply given by 
\[ \ln\left(\frac{c}{1-c}\right), \]
and its standard error can be approximated by applying the \textit{delta method}\(^{29}\). This then yields\(^{14}\):

\[
\text{SE (logit(}c\text{))} \approx \frac{\text{SE}(c)}{c(1-c)} \approx \sqrt{\frac{1 + n^* \frac{1-c}{2-c} + m^* c}{mnc(1-c)}} \quad (4)
\]

Alternatively, when the lower \((c_{lb})\) and upper \((c_{ub})\) boundary of the confidence interval of the \( c \)-statistic are available, it is preferred to derive the standard error of the logit \( c \)-statistic as follows (to preserve possible asymmetry around the \( c \)-statistic)\(^{30}\):

\[
\text{SE (logit(}c\text{))} \approx \frac{\logit(c_{ub}) - \logit(c_{lb})}{2z^*} \quad (5)
\]

where \( z^* \) is the \( 100(1 - \alpha/2) \) percentile of the Normal distribution. For instance, for a 95% confidence interval we have \( z^* = 1.96 \).

\textbf{EuroSCORE II} Estimates for the \( c \)-statistic could be obtained for all 23 validations (Supporting Information 1.1). For 4 validations, equation 4 was used to approximate the standard error of the (logit) \( c \)-statistic. An example is given below.

Previously, Howell \textit{et al.} assessed the predictive performance of EuroSCORE II in 933 high-risk patients\(^{31}\). The observed in-hospital mortality was 9.7% (90 deaths). The reported \( c \)-statistic was 0.67, however, no information was provided on the corresponding standard error or confidence interval. We can derive the logit \( c \)-statistic as follows:
Debray, Damen, Riley, et al.

$$\text{logit}(c) = \ln \left( \frac{0.67}{1 - 0.67} \right) = 0.708$$

Further, based on the reported information, we have $n = 90$, $m = 843$ and $m^* = n^* = 465.5$. This enables us to estimate the standard error of the logit $c$-statistic (equation 4):

$$\text{SE}(\text{logit}(c)) \approx \sqrt{1 + \frac{465.5 \times \frac{1-0.67}{2-0.67} + \frac{465.5 \times 0.67}{1+0.67}}{90 \times 843 \times 0.67 \times (1 - 0.67)} \approx 0.134$$

**Framingham** The $c$-statistic was only reported in 19/24 validations. In some cases, missing $c$-statistics could be restored by contacting the study authors (2 validations). For these 21 estimates of the $c$-statistic, 10 studies provided the standard error and 11 studies required approximation of the standard error using Newcombe’s method (Supporting Information 1.2).

**Model calibration**

Calibration refers to a model’s accuracy of predicted probabilities, and is preferably reported graphically in so-called calibration plots. In these plots, expected outcome probabilities from the model are depicted against observed outcome frequencies in the validation dataset, often across tenths of predicted risk or for 4-5 risk groups over time (e.g. via Kaplan-Meier plots versus predicted survival). Calibration plots can also be constructed using smoothed loess curves, by directly regressing (transformations of) expected versus observed outcomes.

In order to summarize a model’s calibration performance across different validation studies, it is helpful to retrieve expected and observed outcome probabilities (e.g. across different risk strata), and to extract reported calibration measures. For prediction models for binary outcomes as well as for time-to-event outcomes, common measures are the calibration intercept and slope$^{32}$. The intercept is also known as calibration-in-the-large, and indicates whether predicted probabilities are, on average, too high (intercept below 0) or too low (intercept above 0). Conversely, the calibration slope quantifies whether predicted risks are, on average, too extreme (slope below 1) or too invariant (slope above 1)$^{33}$.

Unfortunately, extraction of calibration measures is often hampered by poor assessment and reporting$^{11,16}$. For instance, validation studies rarely present information on the calibration intercept and slope, or different studies report report estimates for different risk strata or time horizons. However, it is common for validation studies to present information on the total number of observed (O) and expected (E) events, or the corresponding observed frequencies $P_O$ and $P_E$. These quantities can then be used
to calculate the total O:E ratio, which provides a rough indication of the overall model calibration. In particular, the total O:E ratio quantifies the averaged observed:expected ratio across the entire range of predicted risks, and is strongly related to calibration-in-the-large\textsuperscript{14}. Formula’s for calculating the total O:E ratio and its standard error were recently presented (see also Supporting Information 3)\textsuperscript{14}.

For prediction models with a time-to-event outcome, it is important to be aware that some events are likely to be unobserved (e.g. due to drop-out). For this reason, extracted values for $O$ (or $P_O$) should account for censoring (and thus be based on Kaplan-Meier estimates) as otherwise they cannot directly be compared to $E$ (or $P_E$). Additional difficulties may arise when event rates are only available for shorter (or longer) follow-up times than intended for the review aim, as the applicability of calculated O:E ratios may then be limited. This situation may, for instance, arise when a model developed to predict 10 year risk has been validated for 5 year risk. Although it is possible to exclude such studies from the meta-analysis, an alternative strategy that doesn’t discard information is to apply extrapolation by assuming a Poisson distribution (Supporting Information 3.2). When missing, the standard error of the total O:E ratio can directly be derived from the total number of observed events.

Further, as we will discuss in the next section, estimates of the total O:E ratio and its standard error should be transformed to the (natural) log scale prior to meta-analysis\textsuperscript{28}. Again, this can be achieved by applying the delta method\textsuperscript{34} or by transforming reported confidence intervals\textsuperscript{30}. The former strategy generally yields the following approximation\textsuperscript{14}:

$$\text{SE} (\ln(O:E)) \approx \frac{\text{SE}(O)}{O} \approx \sqrt{1 - \frac{P_O}{O}}$$

Table 2 provides an overview of formula’s for calculating the total log O:E ratio and its standard error for prediction models with a time-to-event outcome. Examples are provided in Supporting Information 3.3 and in our recent publication\textsuperscript{14}.

Although the total O:E ratio is a useful measure to depict a model’s overall calibration, it is rarely sufficient to identify whether predicted risks are sufficiently accurate. In particular, substantial miscalibration may still occur even when O:E = 1. For this reason, it is generally recommended to assess calibration performance separately for different subgroups. An example is given by a recent review on the performance of the additive EuroSCORE, where the total O:E ratio was extracted for each distinct EuroSCORE value\textsuperscript{14}. These estimates can then be used to construct a (summary) calibration plot, which is more informative than a single summary estimate of the total O:E ratio. Alternatively, when validation studies report (observed and predicted) event rates across different risk strata, it is possible to estimate a calibration slope of the observed-expected plot\textsuperscript{33}. A poor calibration slope usually reflects overfitting of the model in the development sample, but may also indicate lack of transportability\textsuperscript{9}. The calibration slope therefore may provide further insight into the potential generalizability of the model under review.
Rather than estimating the slope for individual validations, we discuss in the next section how reported event rates across the primary studies can directly be used to obtain a summary estimate of the calibration slope and its standard error.

**Example** Buitrago *et al.* analyzed the 10-year performance of the original Framingham coronary risk functions in nondiabetic patients. They reported the observed ($O_{t=10} = 0.109$) and predicted ($E_{t=10} = 0.169$) 10-year risk estimates in a table, as well as the total number of observed events ($O_{t=10} = 22$). Hence, we have:

$$\ln(O:E)_{t=10} = \ln(0.109) - \ln(0.169) = -0.439$$

with

$$\text{SE} (\ln(O:E)_{t=10}) = \sqrt{(1 - 0.109)/22} = 0.201$$

**Framingham** Estimates for observed and expected 10 year CHD risk were directly available for 6/24 validations (Supporting Information 1.2). For some studies, O:E ratios were derived from reported calibration plots, see for instance example 4 in the supporting information. For 10/24 validations, information was only available for 5 or 7.5 years follow-up, and we therefore considered a sensitivity analysis where we extrapolated observed and expected event counts using a Poisson distribution. A total of 11 validations presented event rates for different risk strata (see, for instance, Figure S5). The corresponding calibration plots are depicted in Figure S2 and S3, and can be used to derive summary estimates of the calibration slope.

**EuroSCORE II** The total number of observed and expected events were available for all 23 validations. Furthermore, because the EuroSCORE II model does not consider time-to-event, estimates for the total O:E ratios and their standard error could directly be derived from extracted quantities for $O$ and $E$ (Supporting Information 1.1).

**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 to provide an overall summary of their performance. We here consider the situation where $K$ studies ($i = 1, \ldots, K$) are available for meta-analysis and describe various meta-analysis approaches or models that can be implemented using a frequentist or a Bayesian estimation framework. All meta-analysis approaches or
models have been implemented in the R package metamisc, which in turn makes use of the metafor package\textsuperscript{36} and the JAGS software\textsuperscript{37} to summarize (restored) estimates of model performance. Relevant examples and source code are available in Supporting Information 4.3.

Meta-analysis models

In general, two main types of meta-analysis models can be distinguished. In fixed effect meta-analysis models, all studies are considered to be equivalent, and variation in predictive performance measures across studies are assumed to arise by chance only. Accordingly, precision estimates of discrimination and calibration parameters are used to weight each study in the averaging of the model’s corresponding performance measures. In random effects meta-analysis models, it is assumed that variation in predictive performance measures across studies may not only appear due to chance, but is also prone to the (potential) presence of between-study heterogeneity. As a result, random effects models usually yield larger confidence intervals and assigns study weights that are more similar to one another than those under fixed effect meta-analysis\textsuperscript{38}.

Although both types of aforementioned meta-analysis models have limitations, and following the guidance of meta-analysis of interventions and diagnostic tests, we generally recommend the use of random effects models. In particular, discrimination and calibration performance are highly dependent on patient spectrum (case-mix variation) and therefore most likely to vary across validation studies\textsuperscript{14}. For instance, it is well known that a model’s discrimination performance tends to deteriorate when it is applied to populations or subgroups with a more homogeneous case-mix, as there is less separation of predicted risks across individuals\textsuperscript{9;10;20;39;40}. Between-study heterogeneity may also appear when reported performance estimates (such as $c$-statistics) are based on different definitions or estimation methods (e.g. adopt different criteria to account for ties or censoring).

Model Discrimination For random effects meta-analysis of the $c$-statistic (for either logistic or survival models) we have:

$$\logit(c_i) \sim \mathcal{N}(\mu_{\text{discr}}, \text{Var}(\logit(c_i)) + \tau^2_{\text{discr}})$$

(Model 1)

with $\logit(c_i)$ the logit of the estimated $c$-statistic in the $i$\textsuperscript{th} study. The logit transformation is applied to improve the validity of the Normality assumption, which is used to model variability around the summary $c$-statistic\textsuperscript{28}. After estimation of Model 1, the summary $c$-statistic is given by $\logit^{-1}(\hat{\mu}_{\text{discr}})$, which corresponds to $1/(1 + \exp(-\hat{\mu}_{\text{discr}}))$. The extent of between-study heterogeneity is quantified by $\tau_{\text{discr}}$, the between-study standard deviation of the logit $c$-statistic.
**Model Calibration** Similarly, for random effects meta-analysis of the total O:E ratio, as a measure of model calibration, we have:

$$\ln(O:E)_i \sim \mathcal{N}(\mu_{\text{cal.OE}}, \text{Var}(\ln(O:E)_i) + \tau_{\text{cal.OE}}^2)$$  \hspace{1cm} \text{(Model 2)}$$

with $\ln(O:E)_i$ the natural log of the estimated O:E ratio in the $i^{th}$ study, and $\text{Var}(\ln(O:E)_i)$ its error variance. Again, a transformation is applied to improve the validity of the Normality assumption\textsuperscript{28}. Here, the summary O:E ratio is given by $\exp(\hat{\mu}_{\text{cal.OE}})$. An important limitation of Model 2 is that continuity corrections are required when no events were observed or expected. This situation may, for instance, arise when validation studies are relatively small or when their event rate is much lower than in the development study.

One approach to avoid continuity corrections is to explicitly account for sampling error in the observed event rates by modelling the binomial likelihood directly\textsuperscript{41}:

$$O_i \sim \text{Binom}(N_i, p_{O,i})$$

$$\ln\left(\frac{p_{O,i}}{p_{E,i}}\right) \sim \mathcal{N}(\mu_{\text{cal.OE}}, \tau_{\text{cal.OE}}^2)$$  \hspace{1cm} \text{(Model 2*)}$$

Alternatively, we may adopt a Poisson-Normal model and model the data as counts:\textsuperscript{41}:

$$O_i \sim \text{Poisson}(E_i \exp(\eta_i))$$

$$\eta_i \sim \mathcal{N}(\mu_{\text{cal.OE}}, \tau_{\text{cal.OE}}^2)$$  \hspace{1cm} \text{(Model 2**)}$$

where the summary O:E ratio is again given by $\exp(\hat{\mu}_{\text{cal.OE}})$. Unfortunately, Model 2* and Model 2** may no longer be advantageous when some studies do not report $E_i$ (or $p_{E,i}$) but provide estimates of the total O:E ratio and its standard error. In such situations, continuity corrections can be avoided more effectively by considering study-specific likelihood functions. This approach is described in more detail below, and in Supporting Information 4.3.2.

Finally, if observed and expected event rates are available for different strata of predicted risk in the validation studies, it is possible to obtain a summary estimate of the calibration slope (Supporting Information 4.3.3). The model below is a natural extension of Cox’ proposed regression model for describing agreement between predicted and observed probabilities\textsuperscript{33}.
\[ O_{ij} \sim \text{Binom}(N_{ij}, p_{O,ij}) \]
\[ \logit(p_{O,ij}) = \alpha_i + \beta_i \logit(P_{E,ij}) \quad \text{(Model 3)} \]
\[ \beta_i \sim \mathcal{N}(\mu_{\text{cal.slope}}, \tau_{\text{cal.slope}}^2) \]

with \( O_{ij} \) the number of observed events in subgroup \( j \) of study \( i \), which is modeled according to a binomial distribution with event probability \( p_{O,ij} \). For prediction models with a survival outcome, \( O_{ij} \) and \( N_{ij} \) represent quantities for specific (and same) time periods, and should be adjusted when drop-out occurred or when studies evaluated different time periods (e.g. by means of extrapolation). The summary estimate of the calibration slope is simply given by \( \hat{\mu}_{\text{cal.slope}} \).

**The frequentist approach for random effects meta-analysis**

The parameters of the meta-analysis models above (Model 1 and Model 2) can be estimated by optimizing their corresponding (log-)likelihood function. This yields the well-known estimator of the meta-analysis summary\(^{42}:\)

\[
\hat{\mu} = \frac{\sum_{i=1}^{K} \left( \hat{\theta}_i / (\tau^2 + \text{Var}(\theta_i)) \right)}{\sum_{i=1}^{K} (1 / (\tau^2 + \text{Var}(\theta_i)))}
\]
\[
\text{SE}(\hat{\mu}) = \sqrt{\frac{1}{\sum_{i=1}^{K} (1 / (\tau^2 + \text{Var}(\theta_i)))}}
\]

where \( \hat{\theta}_i \) represent the study specific estimates for the parameter of interest (e.g. the logit \( c \)-statistic or log O:E ratio). In the standard DerSimonian and Laird approach, the heterogeneity parameter \( \tau^2 \) is estimated separately and subsequently inserted in the equations above\(^{42,43}.\) We recommend to use restricted maximum likelihood (REML) estimation to account for the simultaneous estimation of \( \tau \) and \( \mu \). Note that for Model 2*, Model 2** and Model 3, estimates for \( \mu \) and \( \text{SE}(\mu) \) are analytically intractable and can only be derived jointly through iterative estimation procedures.

In line with previous recommendations\(^ {44,45}.\), we propose to correct estimates of \( \text{SE}(\hat{\mu}) \) for potential bias when few studies are included in the meta-analysis. This can, for instance, be achieved using the method proposed by Hartung and Knapp:

*Prepared using sagej.cls*
\[ \text{SE}_{\text{HK}}(\hat{\mu}) = \text{SE}(\hat{\mu}) \sqrt{\frac{1}{K - 1} \sum_{i=1}^{K} \left( \frac{\hat{\theta}_i - \hat{\mu}}{\text{Var}(\hat{\theta}_i) + \hat{\tau}^2} \right)^2} \]

For all models, boundaries of the confidence interval (CI) can then be approximated using:

\[ \hat{\mu} \pm t_{K-1} \text{SE}_{\text{HK}}(\hat{\mu}) \]

with \( t_{K-1} \) denoting the 100\((1 - \alpha/2)\) percentile of the Student-\(t\) distribution with \( K - 1 \) degrees of freedom, where \( \alpha \) is usually chosen as 0.05, to give a 5% significance level and thus 95% confidence interval. The Student-\(t\) (rather than the Normal) distribution is chosen to account for the uncertainty in \( \tau^2 \).

Finally, for meta-analysis of model discrimination and calibration, we recommend the calculation of an (approximate) 95% prediction interval (PI) to depict the extent of between-study heterogeneity \(^{46}\). This interval provides a range for the predicted model performance in a new validation of the model. A 95% PI for the summary estimate in a new setting is approximately given as

\[ \hat{\mu} \pm t_{K-2} \sqrt{\hat{\tau}^2 + \left( \text{SE}_{\text{HK}}(\hat{\mu}) \right)^2} \]

where again the Student-\(t\) (rather than the Normal) distribution is used to help account for the uncertainty of \( \hat{\tau} \). Recently, Partlett and Riley showed that this equation has poor coverage in many situations, and requires improvement \(^{44}\). This motivates the Bayesian approach below.

**The Bayesian approach for random effects meta-analysis**

Generally speaking, in contrast to frequentist methods, Bayesian methods use formal probability models to express uncertainty about parameter values \(^{47,48}\). This is particularly relevant when confronting sparse data, multiple comparisons, collinearity, or non-identification, and for deriving PI \(^{49}\). Estimation problems are likely to occur in frequentist meta-analyses of prediction model performance, as validations of a specific model and thus existing evidence or data on predictive performance measures are often sparse. Furthermore, many frequentist estimation methods (including REML) sometimes fail to produce reliable confidence or prediction intervals. For instance, Partlett and Riley showed that the coverage of Hartung-Knapp confidence intervals based on REML estimation are too narrow for meta-analyses with less than
However, when the heterogeneity is small and the study sizes are mixed, the Hartung-Knapp method produces confidence intervals that are too wide. Similarly, prediction intervals have poor coverage when the extent of heterogeneity is limited, when few (less than 5) studies are included, or when there are a mixture of large and small studies. For this reason, several authors have recommended the adoption of a Bayesian estimation framework.

An additional advantage of Bayesian meta-analysis models is that the distribution for modeling the within-study variation can be tailored to each validation study. This is, for instance, relevant when meta-analyzing the total O:E ratio. In particular, a binomial distribution can be used for studies that report the total number of observed events, the total number of expected events and the total sample size (Model 2*). Conversely, a Poisson distribution can be used for studies that only report the total number of observed and expected events (Model 2**). Finally, a Normal distribution can be used for studies providing information on the total O:E ratio and its standard error (Model 2). The corresponding likelihood functions are then linked together by constructing a shared parameter model through $\mu_{\text{cal훼}}$ and $\tau_{\text{cal OE}}$ (Supporting Information 4.3.2).

When adopting a Bayesian framework for meta-analysis, it is always important to specify appropriate prior distributions. In aforementioned meta-analysis models, prior distributions are needed for the unknown parameters $\mu_{\text{discr}}$, $\mu_{\text{cal OE}}$, $\mu_{\text{cal slope}}$, $\alpha_i$, $P_{E,ij}$, $\tau_{\text{discr}}$, $\tau_{\text{cal OE}}$ and $\tau_{\text{cal slope}}$. It is generally recommended to be conservative when setting the spread of the prior, and to empirically justify prior assertions. For this reason, we here propose the use of non-informative prior distributions by assuming that $\mu_{\text{discr}}$, $\mu_{\text{cal slope}}$, $\alpha_i$, $\logit(P_{E,ij}) \sim \mathcal{N}(0, 10^6)$. Further, because $\mu_{\text{cal OE}}$ requires exponentiation and is often based on discrete likelihood functions, we here assume that $\mu_{\text{cal OE}} \sim \mathcal{N}(0, 100)$. In line with previous recommendations, we consider empirically based priors for $\tau_{\text{discr}}$, $\tau_{\text{cal OE}}$ and $\tau_{\text{cal slope}}$ to rule out unreasonable values for these parameters and to allow for regularization when few studies are available for meta-analysis. Therefore, we retrieved heterogeneity estimates from more than 20 previously published meta-analyses (Supporting Information 4.1). Finally, we assume prior independence of $\mu$ and $\tau$.

Results in Supporting Information 4.2 demonstrate that for the logit c-statistic and the log O:E ratio, it is unlikely that $\tau_{\text{discr}} > 2$ and, respectively, $\tau_{\text{cal OE}} > 2$. Although no empirical data was identified for heterogeneity on the calibration slope, individual estimates usually do not exceed [0,2] and thus it is also unlikely that $\tau_{\text{cal slope}} > 2$. For this reason, we propose the uniform distribution as functional form for modeling the prior, with $\tau_{\text{discr}}$, $\tau_{\text{cal OE}}$, $\tau_{\text{cal slope}} \sim \text{Unif}(0, 2)$.

Because the uniform distribution tends to unduly favor presence of heterogeneity in discrimination and calibration estimates across studies, we also consider a half Student-t distribution with location $m$, scale $\sigma$ and $\nu$ degrees of freedom. Here, we set $m = 0$ and define $\sigma$ equal to the largest empirical value of $\hat{\tau}$ (to allow for more extreme values of heterogeneity than presented in

Prepared using sagej.cls
Supporting Information 4.1). These hyper-parameter values allow to penalize the extent of between-study heterogeneity when the number of included validation studies is low. Further, we chose \( \nu = 3 \) to ensure that the variance \( \sigma^2 \nu / (\nu - 2) \) exists. Finally, we truncate samples for \( \tau \) above 10 to rule out unreasonable values. The resulting priors are then given as \( \tau_{\text{discr}} \sim \text{Student-}t(0, 0.5^2, 3) \, T[0, 10] \) and \( \tau_{\text{cal.OE}} \sim \text{Student-}t(0, 1.5^2, 3) \, T[0, 10] \) and \( \tau_{\text{cal.slope}} \sim \text{Student-}t(0, 0.5^2, 3) \, T[0, 10] \). Results in Supporting Information 4.3 indicate that aforementioned priors allow for large, but realistic, values of between-study heterogeneity.

To evaluate the performance of the proposed prior distributions, we used the empirical example data to conduct a series of simulation studies where we varied the size of the meta-analysis (Supporting Information 4.4). Results in Figure 1 suggest that the proposed priors substantially improve estimation of \( \tau_{\text{discr}} \) when meta-analyses are relatively small. Researchers may, however, need to further tailor these priors to their own particular setting and example.

To further highlight the potential advantages of Bayesian meta-analysis, we extended Model 3 to directly account for (1) sampling error in the number of expected events (2) uncertainty due to the need for extrapolation and (3) uncertainty resulting from data transformations. The corresponding Model 3* is given in Supporting Information 4.3.3, and may help to conduct sensitivity analyses.

Finally, similar to a frequentist meta-analysis, uncertainty around summary estimates resulting from sampling error and/or heterogeneity can be formally quantified. In a Bayesian meta-analysis, the corresponding intervals are denoted as credibility and, respectively, prediction intervals. Boundaries for these intervals can directly be obtained by sampling from the posterior distribution. It is also possible to make probabilistic statements about future validation studies, such as the probability that the \( c \)-statistic will be above 0.7, or that the calibration slope will be between 0.9 and 1.1.

**Empirical examples**

All meta-analysis models have been implemented in the R-package `metamisc`. Convergence of Bayesian meta-analysis models was verified by evaluating the upper confidence limits of the potential scale reduction factor (values \( \leq 1.05 \) were considered as converged).

**EuroSCORE II** Results of the frequentist meta-analysis are depicted in Figure 3. For model discrimination, we found a summary \( c \)-statistic of 0.79 (95% CI 0.76 to 0.81; approximate 95% PI 0.68 to 0.87). For model calibration, we found a summary O:E ratio of 1.11 (95% CI 0.90 to 1.36; approximate 95% PI 0.43 to 2.85). Although similar results were obtained when implementing alternative link functions, prediction intervals for the summary O:E ratio were notably larger when adopting a Bayesian estimation framework (Table 3). Overall, the results suggest that, as expected, discrimination, and particularly calibration performance of EuroSCORE II substantially varied across studies. Although
the average O:E ratio was close to 1, predicted risks were systematically too high or too low in certain populations as revealed by the PI. The calibration slope could not be estimated for EuroSCORE II because subgroup-specific event rates were not extracted from the validation studies and the slope itself was not directly provided.

**Framingham** Results in Table 4 reveal substantial heterogeneity for the discrimination performance, with 95% PI ranging from 0.59 to 0.77 based on the frequentist approach. Although the 10-year predicted risk of CHD occurrence was usually too large (summary O:E = 0.6), the calculated PI (0.2 to 1.8) indicates that predicted risks were too low for some of the validation studies. The summary estimate of the calibration slope was close to 1, indicating that predictions by the Framingham Risk Score were sufficiently scaled (despite being too high on average). Further, because most studies only reported calibration performance for 5 years follow-up, we conducted two sensitivity analyses where (1) studies with less than 10 years follow-up were excluded from the meta-analysis, and (2) where observed and expected event counts were extrapolated using a Poisson distribution. Results in Supporting Information 5.1 indicate that both strategies yielded similar summary estimates.

**Investigating sources of heterogeneity**

As applies to all types of meta-analysis, also for meta-analysis of prediction model studies, summary estimates of model performance may be of limited value in the presence of (substantial) between-study heterogeneity. Although we recommended the use of random effects models to evaluate the presence of heterogeneity, these models do not offer any insight about potential causes of this. For this reason, it is often helpful to investigate potential sources of heterogeneity in the predictive performance by performing a meta-regression or subgroup analysis\textsuperscript{14,60}. Common sources of heterogeneity are differences in study characteristics\textsuperscript{61}, differences in study quality, or differences in baseline risk across the validation studies\textsuperscript{9,10}. Heterogeneity may also arise when reported performance estimates are based on “improper” validations where certain predictors were neglected (e.g. due to missing data) or where various model parameters have been adjusted (e.g. intercept update).

**Meta-regression models**

We extend the presented meta-analysis models as follows to investigate sources of between-study heterogeneity\textsuperscript{62}. Let $X_i$ denote the (row)vector of covariates of study $i$ including the constant term. Meta-regression considers the relation between the “predictor” $X_i\gamma$ and the performance estimate of the corresponding study. Hereby, it assumes that covariate effects $\gamma$ are (roughly) linear on the meta-analysis scale. Similar to aforementioned random effects models, Normal distributions are used to account for
the presence of sampling error and between-study heterogeneity. For instance, the marginal model for meta-regression of the logit $c$-statistic is given by:

$$\text{logit}(c_i) \sim \mathcal{N}(X_i \gamma, \text{Var} (\text{logit}(c_i)) + \tau_{\text{discr}}^2)$$  \quad \text{(Model 1')}$$

To enhance estimation of covariate effects $\gamma$, it is recommended to center the covariates $X_i$ around their respective means. Although meta-regression may help to identify sources of heterogeneity and poor model transportability, the power for this is often low. In addition, meta-regression may suffer from study-level confounding (also known as ecological bias) when analyzing aggregate patient (average) information.

**Empirical examples**

We used the R-package *metafor* to estimate the meta-regression models.

**EuroSCORE II** In a meta-regression model (for discrimination), we examined whether heterogeneity was explained by one or more of the following: the spread of the prognostic index of EuroSCORE II in each validation study, whether the study was a multicentre study, whether the study included patients before 2010 (i.e., before EuroSCORE II was developed) and the spread of the age of the patients. The resulting meta-regression plots are depicted in Figure S9 (Supporting Information), and indicate that the summary $c$-statistic generally remains unaffected by changes in the distribution of the prognostic index or patient age. The p-value of the regression coefficients were all larger than 0.05. We therefore could not attribute heterogeneity in the $c$-statistic to these differences.

**Framingham** We examined whether heterogeneity for the Framingham Risk Score in male populations could be explained by differences in age distribution, smoking prevalence, and observed 10 year event rates. For instance, the mean age ranged from 41 to 74 years across the included validations. Also the standard deviation of age substantially varied across the included validation studies (range: 2 to 13), further highlighting the presence of differences in patient spectrum. Results in Figure 4 indicate that the $c$-statistic of the Framingham Risk Score tended to decrease when validated in older (male) populations ($p=0.01$). This can be explained by the narrower case-mix leading to more narrow separation, and thus lower $c$ values. For the calibration performance, we found that under-estimation (i.e. $O:E < 1$) notably occurred in low-risk populations ($p < 0.01$, Figure 4).
Conclusion

Quantitative synthesis of prediction model studies may help to better understand their potential generalizability and can be achieved by applying meta-analysis methods. In this paper, we discussed several common stumbling blocks when meta-analyzing the predictive performance of prediction models. In particular, substantial efforts are often needed to restore missing information from the primary studies and to harmonize the extracted performance statistics. In addition, estimates of a model’s predictive performance are likely to be affected by the presence of between-study heterogeneity. Finally, because validation studies of a certain prediction model are often sparse, traditional (frequentist) meta-analysis models may suffer from convergence issues and yield unreliable estimates of precision and between-study heterogeneity. For this reason, we presented several methods to overcome these deficiencies, and to obtain relevant summary statistics of prediction model performance even when the primary validation studies did not report such information. Furthermore, to facilitate the implementation of the presented methods, we created the open source R package `metamisc` which includes the empirical example data.

We generally recommend the use of one-stage methods for summarizing the O:E ratio (here, Model 2* and Model 2**) and calibration slope (here, Model 3), as they use the exact likelihood. Further, adopting a Bayesian estimation framework may help to fully propagate the uncertainty resulting from estimating the within-study standard errors and the between-study standard deviation, which in turn may improve the coverage of calculated intervals. Future efforts should focus on comparing the presented meta-analysis methods in new empirical examples and simulation studies. Results from previous studies suggest that these methods are most likely to yield discordance when there is low between-study heterogeneity, when there are few studies for meta-analysis, or when studies are small or based on different sample sizes.

Further work is also still needed to summarize the evidence on multiple, competing, prediction models that were compared head-to-head in validation studies. Another important issue arises when individual participant data (IPD) are available for one or more validation studies. Although it is possible to reduce these studies to relevant performance estimates and to adopt the meta-analysis methods presented in this article, more advanced (so-called one-stage) approaches exist to directly combine the IPD with published summary data. An additional advantage of IPD is that it becomes possible to adjust discrimination performance for (variation in) subject-level covariates, and thus to gain more understanding in sources of between-study heterogeneity. Finally, methods for meta-analyzing clinical measures of performance such as net benefit deserve further attention.

In summary, our empirical examples demonstrate that meta-analysis of prediction models is a feasible strategy despite the complex nature of corresponding studies. As developed prediction models are...
being validated increasingly often, and as the reporting quality is steadily improving, we anticipate that evidence synthesis of prediction model studies will become more commonplace in the near future. The R package *metamisc* is designed to facilitate this endeavor, and will be updated as new methods become available.

**Funding**

The work leading to these results has received support from the Netherlands Organisation for Health Research and Development (91617050) and from the Cochrane Methods Innovation Funds Round 2 (MTH001F).

**Acknowledgements**

We gratefully acknowledge Ian White for sharing the algebraic expressions relating the $c$-statistic in terms of the standard deviation of the linear predictor (Supporting Information 2.2). Further, we thank Hans van Houwelingen for his reference to earlier work by D. R. Cox on the calibration slope.

**References**


<table>
<thead>
<tr>
<th></th>
<th>EuroSCORE II</th>
<th>Framingham Risk Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td>Patients undergoing coronary artery bypass grafting</td>
<td>General male population, free from CHD and not on treatment</td>
</tr>
<tr>
<td><strong>Intervention</strong></td>
<td>EuroSCORE II</td>
<td>Framingham Wilson</td>
</tr>
<tr>
<td><strong>Comparator</strong></td>
<td>Not applicable</td>
<td>Not applicable</td>
</tr>
<tr>
<td><strong>Outcome</strong></td>
<td>All-cause mortality</td>
<td>Fatal or non-fatal coronary heart disease (CHD)</td>
</tr>
<tr>
<td><strong>Timing</strong></td>
<td>Death within 30 days of operation or within the same hospital admission, predicted using preoperative conditions.</td>
<td>Initial CHD within 10 years</td>
</tr>
<tr>
<td><strong>Setting</strong></td>
<td>To inform considerations for the timing and choice of surgical intervention</td>
<td>Target individuals for the primary prevention of coronary disease</td>
</tr>
</tbody>
</table>
### Table 2. Formulas for estimating the total O:E ratio from other information in a primary study

<table>
<thead>
<tr>
<th>What is reported?</th>
<th>Estimate for (O:E)</th>
<th>Estimate for ( \text{Var}(O:E) )</th>
<th>Estimate for ( \text{Var}(\ln(O:E)) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( O_t, E_t )</td>
<td>( O_t/E_t )</td>
<td>( O_t/(E_t)^2 )</td>
<td>( 1/O_t )</td>
</tr>
<tr>
<td>( S_{KM,t}, O_t, E_t )</td>
<td>(1 – ( S_{KM,t} ))/(1 – ( S_{E,t} ))</td>
<td>( S_{KM,t}(1 – S_{KM,t})/N_t(1 – S_{E,t})^2 )</td>
<td>( S_{KM,t}/(N_t(1 – S_{KM,t})) )</td>
</tr>
<tr>
<td>( S_{KM,t}, S_{E,t}, N_t )</td>
<td>(1 – ( S_{KM,t} ))/( P_{E,t} )</td>
<td>( S_{KM,t}(1 – S_{KM,t})/N_t(1 – P_{E,t})^2 )</td>
<td>( S_{KM,t}/(N_t(1 – S_{KM,t})) )</td>
</tr>
<tr>
<td>( S_{KM,t}, P_{E,t}, N_t )</td>
<td>(1 – ( S_{KM,t} ))/( P_{E,t} )</td>
<td>( \text{Var}(S_{KM,t})/(1 – S_{E,t})^2 )</td>
<td>( \text{Var}(S_{KM,t})/(1 – S_{KM,t})^2 )</td>
</tr>
<tr>
<td>( S_{KM,t}, S_{E,t}, \text{Var}(S_{KM,t}) )</td>
<td>1 – ( \exp(t \ln(S_{KM,t})) / l )</td>
<td>( t^2 \text{Var}(S_{KM,t}) \exp(2t \ln(S_{KM,t})) / l )</td>
<td>( t^2 \text{Var}(S_{KM,t}) \exp(2t \ln(S_{KM,t})) / l )</td>
</tr>
<tr>
<td>( S_{KM,t}, P_{E,t}, \text{Var}(S_{KM,t}) )</td>
<td>1 – ( \exp(t \ln(1 – P_{E,t})) / l )</td>
<td>( t^2 \text{Var}(S_{KM,t}) \exp(2t \ln(1 – P_{E,t})) / l )</td>
<td>( t^2 \text{Var}(S_{KM,t}) \exp(2t \ln(1 – P_{E,t})) / l )</td>
</tr>
<tr>
<td>( S_{KM,t}, S_{E,t}, N_t )</td>
<td>1 – ( \exp(t \ln(S_{KM,t})) / l )</td>
<td>( t^2 (1 – S_{KM,t}) \exp(2t \ln(S_{KM,t})) / l )</td>
<td>( t^2 (1 – S_{KM,t}) \exp(2t \ln(S_{KM,t})) / l )</td>
</tr>
<tr>
<td>( S_{KM,t}, P_{E,t}, N_t )</td>
<td>1 – ( \exp(t \ln(1 – P_{E,t})) / l )</td>
<td>( t^2 (1 – S_{KM,t}) \exp(2t \ln(1 – P_{E,t})) / l )</td>
<td>( t^2 (1 – S_{KM,t}) \exp(2t \ln(1 – P_{E,t})) / l )</td>
</tr>
</tbody>
</table>

The quantities \( O_t \) and \( E_t \) represent the total number of observed (using Kaplan-Meier estimates) and expected events at time \( t \). The corresponding observed and expected survival probabilities are given by \( S_{KM,t} \) and \( S_{E,t} (= 1 – P_{E,t}) \). When Kaplan-Meier estimates are not available, it is still possible to approximate \( S_{KM,t} \) using the total number of observed events \( O_t \) during time \( t \) and the sample size \( N_t \) or, if drop-out is negligible, the original sample size \( N \).
### Table 3. Meta-analysis estimates for the EuroSCORE II model

<table>
<thead>
<tr>
<th>Performance</th>
<th>Estimation</th>
<th>Model</th>
<th>$K$</th>
<th>Summary</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$-statistic</td>
<td>REML</td>
<td>Model 1</td>
<td>23</td>
<td>0.79</td>
<td>0.76 – 0.81</td>
<td>0.68 – 0.87</td>
</tr>
<tr>
<td></td>
<td>Bayesian†</td>
<td>Model 1</td>
<td>23</td>
<td>0.79</td>
<td>0.76 – 0.81</td>
<td>0.68 – 0.89</td>
</tr>
<tr>
<td></td>
<td>Bayesian‡</td>
<td>Model 1</td>
<td>23</td>
<td>0.79</td>
<td>0.77 – 0.81</td>
<td>0.68 – 0.87</td>
</tr>
<tr>
<td>Total O:E ratio</td>
<td>REML</td>
<td>Model 2</td>
<td>23</td>
<td>1.11</td>
<td>0.90 – 1.36</td>
<td>0.43 – 2.85</td>
</tr>
<tr>
<td></td>
<td>Bayesian†</td>
<td>Model 2*</td>
<td>23</td>
<td>1.10</td>
<td>0.88 – 1.34</td>
<td>0.27 – 2.46</td>
</tr>
<tr>
<td></td>
<td>Bayesian‡</td>
<td>Model 2*</td>
<td>23</td>
<td>1.10</td>
<td>0.88 – 1.33</td>
<td>0.27 – 2.45</td>
</tr>
<tr>
<td></td>
<td>ML*</td>
<td>Model 2**</td>
<td>23</td>
<td>1.09</td>
<td>0.90 – 1.33</td>
<td>0.44 – 2.67</td>
</tr>
<tr>
<td></td>
<td>Bayesian†</td>
<td>Model 2**</td>
<td>23</td>
<td>1.10</td>
<td>0.88 – 1.33</td>
<td>0.29 – 2.45</td>
</tr>
<tr>
<td></td>
<td>Bayesian‡</td>
<td>Model 2**</td>
<td>23</td>
<td>1.10</td>
<td>0.88 – 1.33</td>
<td>0.30 – 2.45</td>
</tr>
</tbody>
</table>

$K = \text{Number of studies included in the meta-analysis; REML = Restricted Maximum Likelihood; ML = Maximum Likelihood; CI = confidence (in case of REML) or credibility (for Bayesian models) interval; PI = (approximate) prediction interval}$

† A uniform prior was used for modeling the between-study standard deviation
‡ A truncated Student-\(t\) distribution was used for modeling the between-study standard deviation

### Table 4. Meta-analysis estimates for the Framingham Risk Score

<table>
<thead>
<tr>
<th>Performance</th>
<th>Estimation</th>
<th>Model</th>
<th>$K$</th>
<th>Summary</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$-statistic</td>
<td>ML</td>
<td>Model 1</td>
<td>21</td>
<td>0.69</td>
<td>0.66 – 0.71</td>
<td>0.59 – 0.77</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>Model 1</td>
<td>21</td>
<td>0.69</td>
<td>0.66 – 0.71</td>
<td>0.59 – 0.77</td>
</tr>
<tr>
<td></td>
<td>Bayesian†</td>
<td>Model 1</td>
<td>21</td>
<td>0.69</td>
<td>0.66 – 0.71</td>
<td>0.59 – 0.78</td>
</tr>
<tr>
<td></td>
<td>Bayesian‡</td>
<td>Model 1</td>
<td>21</td>
<td>0.69</td>
<td>0.66 – 0.71</td>
<td>0.59 – 0.77</td>
</tr>
<tr>
<td>Total O:E ratio</td>
<td>ML</td>
<td>Model 2</td>
<td>17</td>
<td>0.60</td>
<td>0.46 – 0.79</td>
<td>0.20 – 1.84</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>Model 2</td>
<td>17</td>
<td>0.60</td>
<td>0.46 – 0.79</td>
<td>0.19 – 1.90</td>
</tr>
<tr>
<td></td>
<td>Bayesian†</td>
<td>Model 2*</td>
<td>17</td>
<td>0.60</td>
<td>0.44 – 0.78</td>
<td>0.10 – 1.64</td>
</tr>
<tr>
<td></td>
<td>Bayesian‡</td>
<td>Model 2*</td>
<td>17</td>
<td>0.60</td>
<td>0.44 – 0.77</td>
<td>0.10 – 1.61</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>Model 2**</td>
<td>17</td>
<td>0.60</td>
<td>0.46 – 0.78</td>
<td>0.20 – 1.83</td>
</tr>
<tr>
<td></td>
<td>Bayesian†</td>
<td>Model 2**</td>
<td>17</td>
<td>0.60</td>
<td>0.44 – 0.78</td>
<td>0.10 – 1.63</td>
</tr>
<tr>
<td></td>
<td>Bayesian‡</td>
<td>Model 2**</td>
<td>17</td>
<td>0.60</td>
<td>0.44 – 0.78</td>
<td>0.08 – 1.58</td>
</tr>
<tr>
<td>Calibration slope</td>
<td>ML</td>
<td>Model 3</td>
<td>11</td>
<td>1.03</td>
<td>0.94 – 1.12</td>
<td>0.93 – 1.13</td>
</tr>
<tr>
<td></td>
<td>Bayesian†</td>
<td>Model 3</td>
<td>11</td>
<td>1.04</td>
<td>0.87 – 1.21</td>
<td>0.55 – 1.54</td>
</tr>
<tr>
<td></td>
<td>Bayesian‡</td>
<td>Model 3</td>
<td>11</td>
<td>1.04</td>
<td>0.88 – 1.20</td>
<td>0.58 – 1.52</td>
</tr>
</tbody>
</table>

$K = \text{Number of studies included in the meta-analysis; REML = Restricted Maximum Likelihood; ML = Maximum Likelihood; CI = confidence (in case of REML) or credibility (for Bayesian models) interval; PI = (approximate) prediction interval}$

† A uniform prior was used for modeling the between-study standard deviation
‡ A truncated Student-\(t\) distribution was used for modeling the between-study standard deviation
Estimates of $\hat{\tau}_{\text{discr}}$ for meta-analyses with $K$ validation studies. Results are based on $23!/(K!(23-K)!)$ meta-analyses for EuroSCORE II, and on $21!/(K!(21-K)!)$ meta-analyses for the Framingham Risk Score. For Bayesian meta-analysis models, $\hat{\tau}_{\text{discr}}$ represents the posterior median. The reference value (solid line) was obtained by adopting a Bayesian meta-analysis with uniform prior in the full set of 23 (for EuroSCORE II) or 21 (for the Framingham Risk Score) studies. Similar reference values were obtained for ML, REML and Bayes (Student-t).

ML = Maximum Likelihood estimation; REML = Restricted Maximum Likelihood estimation; BAYES (Student-t) = Bayesian estimation with $\tau_{\text{discr}} \sim \text{Student-t}(0, 0.5^2, 3) \times [0, 10]$; BAYES (Uniform) = Bayesian estimation with $\tau_{\text{discr}} \sim \text{Unif}(0, 2)$.
Estimates of $\tau_{cal.OE}$ for meta-analyses with $K$ validation studies. Results are based on $23!/(K!(23 - K)!)$ meta-analyses for EuroSCORE II, and on $17!/(K!(17 - K)!)$ meta-analyses for the Framingham Risk Score. For Bayesian meta-analysis models, $\hat{\tau}_{cal.OE}$ represents the posterior median. The reference value (solid line) was obtained by adopting a Bayesian meta-analysis with uniform prior in the full set of 23 (for EuroSCORE II) or 17 (for the Framingham Risk Score) studies. Similar reference values were obtained for ML, REML and Bayes (Student-t).

ML = Maximum Likelihood estimation; REML = Restricted Maximum Likelihood estimation; BAYES (Student-t) = Bayesian estimation with $\tau_{cal.OE} \sim Student-t(0, 1.5^2, 3) \cdot T[0, 10]$; BAYES (Uniform) = Bayesian estimation with $\tau_{cal.OE} \sim Unif(0, 2)$. 

Prepared using sagej.cls
Figure 3. Meta-analysis estimates for EuroSCORE II.

Figure 4. Results from random-effects meta-regression models for the Framingham Risk Score (validations in male populations). Full lines indicate the bounds of the 95% confidence interval around the regression line. Dots indicate the included validation studies.