Supporting Information

Contents

1 Motivating examples

1.1 The EuroSCORE II model

Table S1: Multivariable regression coefficients of the EuroSCORE II model

Risk factor	**Coefficient**
New York Heart Association class	
II	0.1070545
III	0.2958358
IV	0.5597929
Canadian Cardiovascular Society class 4 angina	0.2226147
Insulin-dependent diabetes mellitus	0.3542749
Age	0.0285181
Female	0.2196434
Extracardiac arteriopathy	0.5360268
Chronic pulmonary dysfunction	0.1886564
N/M mob	0.2407181
Previous cardiac surgery	1.1185990
Renal dysfunction	
On dialysis	0.6421508
Creatinine clearance ≤ 50	0.8592256
Creatinine clearance 50–85	0.3035530
Active endocarditis	0.6194522
Critical	1.0865170
Left ventricle function	
Moderate (31 – 50%)	0.3150652
Poor (21 – 30%)	0.8084096
Very poor (21% or less)	0.9346919
Recent myocardial infarction	0.1528943
Pulmonary artery systolic pressure	
31 – 55 mmHg	0.1788899
≥ 55	0.3491475
Urgency	
Urgent	0.3174673
Emergency	0.7039121
Salvage	1.3629470
Weight of procedure	
1 non-CABG	0.0062118
2	0.5521478
3+	0.9724533
Thoracic aorta	0.6527205
Constant	−5.3245370

N/M mob: neurological or musculoskeletal dysfunction severely affecting mobility; '1 non-CABG': single major cardiac procedure which is not isolated CABG; 2: two major cardiac procedures; 3+: three or more major cardiac procedures. For age, $X_i = 1$ if patient age ≤ 60; X_i increases by one point per year thereafter (age 60 or less $X_i = 1$; age 61 if $X_i = 2$; age 62 if $X_i = 3$ and so on). More information can be found in the original publication [1].

Table S2: Summarised results of the 22 validation studies of the EuroSCORE II model included in our meta-analysis

Study	**Country**	**Enrolment (Years)**	**Number of patients**	**Observed in-hospital mortality**	**Expected in-hospital mortality**	**C statistic**
Nashef*	43 countries	2010	5553	232	219.34	0.8095 (0.014)
Biancari	Finland	2006-2011	1027	28	46.22	0.867 (0.035)
Di Dedda	Italy	2006-2011	1090	41	33.79	0.81 (0.036)
Chalmers	UK	2006-2010	5576	191‡	206.96	0.79 (0.010)
Grant	UK	2010-2011	23740	746	809.53	0.808 (0.008)
Carneo-Alcazar	Spain	2005-2010	3798	215	169.39	0.85 (0.010)
Kunt	Turkey	2004-2012	428	34	7.28	0.72 (0.051)
Kirmani	UK	2001-2010	15497	547	392.07	0.818 (0.007)
Howell	UK/NL	2006-2011	933	90	105.43†	0.67 (0.030)†
Wang (a)	China	2008-2011	11170	226	290.42†	0.72 (0.015)
Borde	India	2011-2012	498	8	10.01	0.72 (0.090)†
Qadir	Pakistan	2006-2010	2004	76	74.55	0.84 (0.023)†
Spiliopoulos	Germany	1999-2005	216	14	8.62	0.77 (0.067)
Wendt	Germany	1999-2012	1066	45	34.11	0.72 (0.034)
Laurent	France	2009-2011	314	18	7.22	0.77 (0.061)
Wang (b)	New Zealand	2010-2012	818	13	13.09†	0.642 (0.071)
Nishida	Japan	1993-2013	461	33	34.11	0.7697 (0.042)†
Barili (a)	Italy	2006-2012	12201	210	305.03	0.8 (0.015)
Barili (b)	Italy	2006-2012	1670	125	103.54	0.82 (0.020)
Paparella	Italy	2011-2012	6191	300	272.40	0.83 (0.012)
Carosella	Argentina	2008-2012	250	9	4.10	0.76 (0.056)
Borracci	Argentina	2012-2013	503	21	16.00	0.856 (0.033)
Osnabrugge	US	2003-2012	50588	1071	1568.23	0.77 (0.010)

* Original development study

† The standard error of the c-statistic was estimated using a method proposed by Newcombe (method 4 in [2])

For the concordance statistic, estimates are presented with corresponding standard error; UK = The United Kingdom; NL = The Netherlands; US = The United States

1.2 The Framingham Risk Score

The Framingham Risk Score was developed in 1998 using data from 2489 men and 2856 women to predict the risk of initial coronary heart disease (CHD) in a free-living population not on medication [3]. The Framingham Risk Score consists of several models that combine information on blood pressure, smoking history, TC and HDL-C levels, diabetes, and left ventricular hypertrophy on the ECG. The regression coefficients of these models are presented in Table S3.

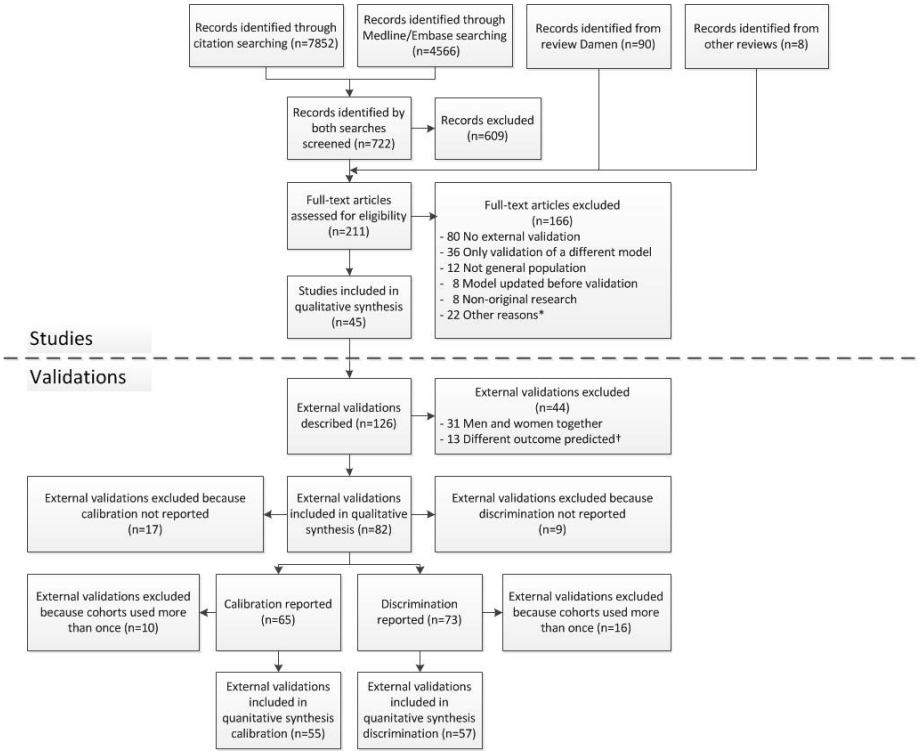
We recently performed a systematic review to assess the performance of the Framingham Risk Score, as well as two other common cardiovascular risk prediction models, in a general population setting [4, 5]. The search identified 820 references, of which 211 were screened in full text (Figure S1). Here, we focus on the studies that assessed the predictive performance of the Framingham Risk Score in male populations. Results from the corresponding 23 validations are presented in Table S4, together with the results from the original development study.

Table S3: Multivariable regression coefficients of the Framingham Risk Score

Risk factor	**Coefficient**		**Coefficient**	
	Men	Women	Men	Women
Age,y	0.04826	0.33766	0.04808	0.33994
Age squared, y		−0.00268		−0.0027
TC, mg/dL				
< 160	−0.65945	−0.26138		
160 – 199	Referent	Referent		
200 – 239	0.17692	0.20771		
240 – 279	0.50539	0.24385		
≥ 280	0.65713	0.53513		
LDL-C, mg/dL				
< 100			−0.69281	−0.42616
100 – 129			Referent	Referent
130 – 159			0.00389	0.01366
160 – 189			0.26755	0.26948
≥ 190			0.56705	0.33251
HDL-C, mg/dL				
< 35	0.49744	0.84312	0.48598	0.88121
35 – 44	0.24310	0.37796	0.21643	0.36312
45 – 49	Referent	0.19785	Referent	0.19247
50 – 59	−0.05107	Referent	−0.04710	Referent
≥ 60	−0.48660	−0.42951	−0.34190	−0.35404
Blood pressure				
Optimal	−0.00226	−0.53363	−0.02642	−0.51204
Normal	Referent	Referent	Referent	Referent
High normal	0.28320	−0.06773	0.30104	−0.03484
Stage I hypertension	0.52168	0.26288	0.55714	0.28533
Stage II–IV hypertension	0.61859	0.46573	0.65107	0.50403
Diabetes	0.42839	0.59626	0.42146	0.61313
Smoker	0.52337	0.29246	0.54377	0.29737
Baseline survival function at 10 years, S(t)	0.90015	0.96246	0.90017	0.9628

The regression coefficients given are used to compute a linear function. The latter is corrected for the averages of the participants' risk factors, and the subsequent result is exponentiated and used to calculate a 10-year probability of CHD after insertion into a survival function. More information, as well as a worked out example, can be found in the original publication [3].

Figure S1: Flow diagram of selected studies



Two searches were performed; one in Medline and Embase and one in Scopus and Web of Science. Studies identified by both searches were screened for eligibility. In the bottom part of the figure, external validations are excluded because cohorts were used more than once to validate the same model.

* E.g. no cardiovascular outcome, not written in English.

† The Wilson and ATPIII model are developed to predict the risk of fatal or nonfatal CHD and the PCE are developed to predict the risk of fatal or nonfatal CVD. External validations that used a different outcome were excluded from the analyses.

Table S4: Summarised results of the included validation studies of the Framingham Risk Score when applied in male populations.

Study	**Country**	**Enrolment (Years)**	**Number of patients**	**Number of events**	**Observed 10y CHD risk**	**Predicted 10y CHD risk**	**C statistic**	
							Est.	SE
Buitrago [6]	Spain	1994-2004	201	22	10.9%	16.9%	0.63	0.059
Comin [7]	Spain	1995-1998	2447	137	7.7%‡°	19.8%‡°	0.679‡	0.023‡
D'Agostino [8]	US	1987-1988	4705	149	6.5%°	7.0%°	0.75	0.020†
D'Agostino [8]	US	1987-1988	1428	46	6.6%°	7.4%°	0.67	0.040†
D'Agostino [8]	US	1982	901	182			0.63	0.023†
D'Agostino [8]	US	1980-1982	2755	77	5.8%°	12.4%°	0.72	0.029†
D'Agostino [8]	Puerto Rico	1965-1968	8713	107	2.6%°	7.3%°	0.69	0.026†
D'Agostino [8]	US	1989-1991	1527	46	6.5%°	9.3%°	0.69	0.039†
D'Agostino [8]	US	1989-1990	956	71			0.63	0.034†
DeFilippis [9]	US	2000-2002	1961	164	8.36%	12.8%	0.69	0.02‡
Empana [10]	Northern Ireland	1991-1993	2399	120	9.8%°	13.4%°	0.66	0.025†
Empana [10]	France	1991-1993	7359	197	5.3%°	12.6%°	0.68	0.019†
Ferrario [11]	Italy	1983-1996	6865	312	5%‡	14.3%‡	0.723	0.028
Jee [12]	South Korea	1996-2001	164005	2086				
Lloyd-Jones [13]	US	1971	2716					
Mainous [14]	US	1987-1989	6239				0.691	0.011
Marrugat [15]	Spain	1995-1998	2447	98	15.4%°	35.6%°	0.68	0.024‡
Reissigova [16]	Czech Republic	1975-1979	646	83	2.4%	10.8%	0.638	0.027
Rodondi [17]	US	1997-1998	981	205	26.6%°	24.5%°	0.583	0.024
Ryckman [18]	US	2004-2005	284	19				
Simmons [19]	United Kingdom	1993-1998	4513	430	9.7%‡	17.7%‡	0.71	0.010
Suka [20]	Japan	1991-1993	5611	80		3.7%	0.71	0.029†
Vaidya [21]	US	1983-1996	404	81	19.8%	11.6%	0.698‡	0.03‡
Wilson* [3]	US	1971-1974	2489	383			0.79	0.013†

* Original development study

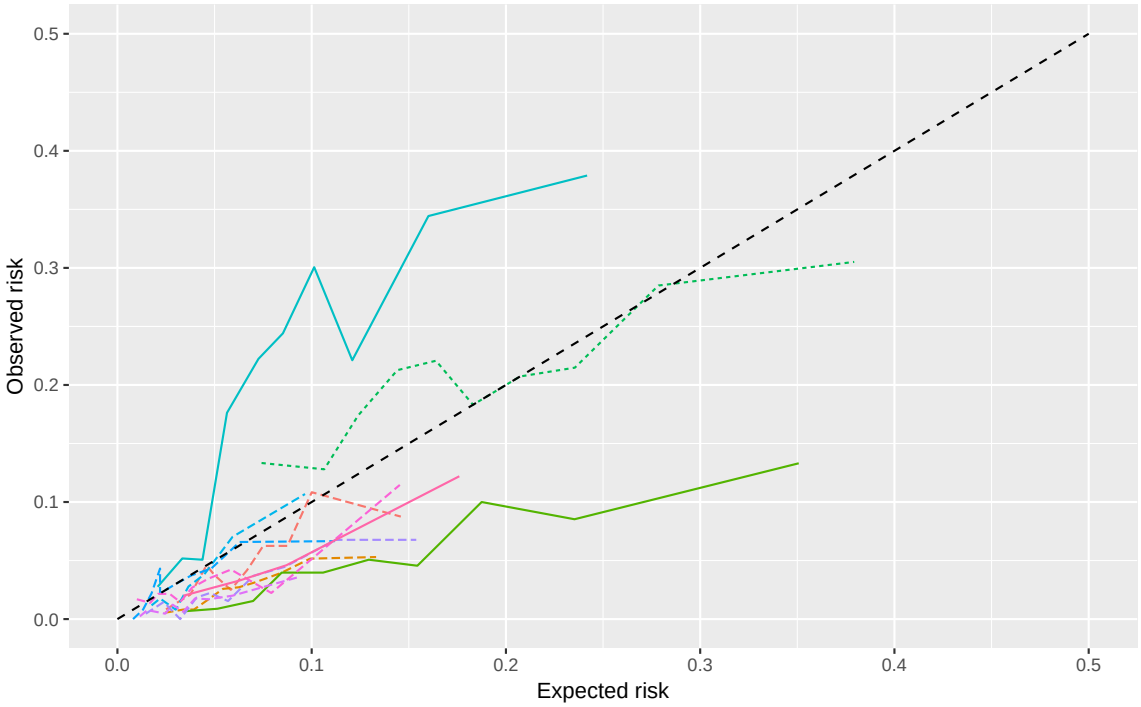
† The standard error of the c-statistic was estimated using a method proposed by Newcombe (method 4 in [2])

‡ Information provided by the study authors

° Risk estimates were extrapolated using the exponential distribution such that 10y survival $S_{10} = \exp(\ln(S_t) \times 10/t)$

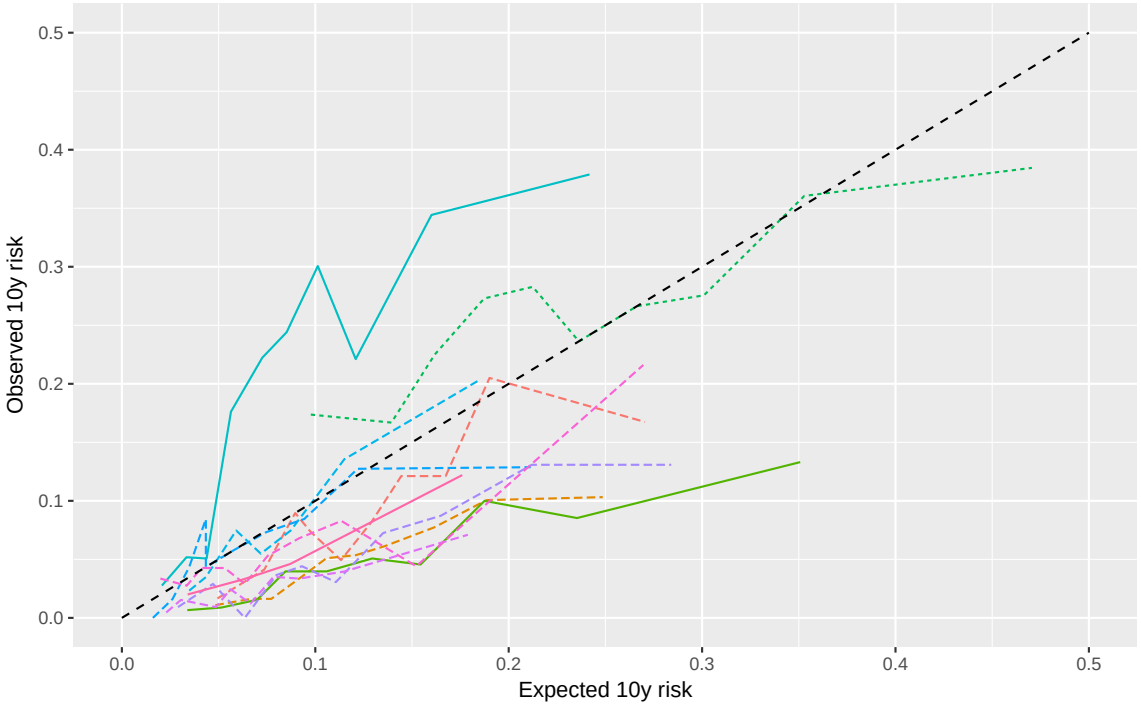
For the concordance statistic, estimates are presented with corresponding standard error; UK = The United Kingdom; NL = The Netherlands; US = The United States

Figure S2: Extracted risk estimates for the Framingham Risk Score when applied to male populations



The diagonal line indicates perfect calibration. Risk estimates were reported for 5 years follow-up (dashed lines), 7.5 years follow-up (dotted lines) and 10 years follow-up (full lines).

Figure S3: Extrapolated risk estimates for the Framingham Risk Score when applied to male populations



The diagonal line indicates perfect calibration. Note that some risk estimates were extrapolated from 5 years (dashed lines) or 7.5 years (dotted lines) to 10 years by assuming a Poisson distribution (see section 3.2).

2 Data preparation of model discrimination

2.1 Estimating the standard error of the c-statistic

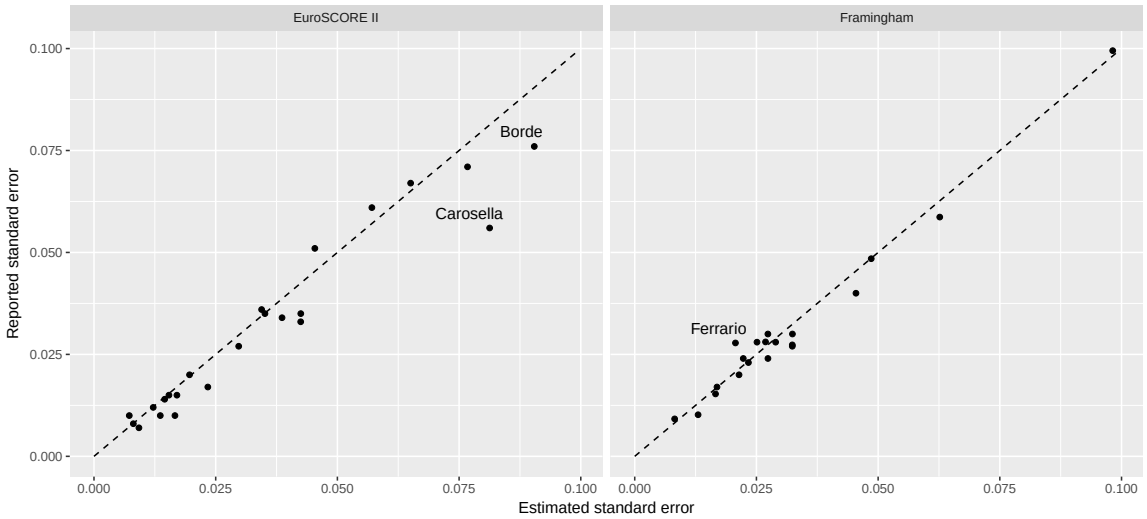
When no standard errors are available for reported c-statistics, they can be approximated using the equations described by Newcombe [2]. We here consider so-called method 4, which is based on a suggestion by Hanley and McNeil [22]. In this method, the error variance of the c-statistic is given by:

$$\widehat{\mathrm{Var}}(\hat{c}) = \frac{\hat{c}\,(1-\hat{c})\left[1 + n^*\frac{1-\hat{c}}{2-\hat{c}} + \frac{m^*\hat{c}}{1+\hat{c}}\right]}{mn} \tag{1}$$

with $\hat{c}$ the estimated c-statistic, $n = O$ (the number of observed events), $m = N - O$ (the total number of non-events) and $m^* = n^* = \frac{1}{2}(m+n) - 1$.

To assess the accuracy of equation 1 in binary and time-to-event data, we estimated the standard error of the c-statistic for all validation studies of EuroSCORE II and Framingham Wilson (male, female and mixed populations) where the *actual* standard error was reported. The resulting discrepancies between estimated and reported standard errors are presented in Figure S4.

Figure S4: Comparison of estimated and reported standard error of the c-statistic in the empirical examples



Estimates for the standard error of the c-statistic were obtained using method 4 proposed by Newcombe.

For the EuroSCORE II model, outliers appear for Carosella [23] and Borde [24]. Both of these studies have fewer than 10 observed events, which might explain the poor accuracy of estimated standard errors. For the Framingham Risk Score, no substantial differences between the estimated and reported standard errors were found. It is, however, important to note that Ferrario [11] adopted a split-sample approach to calculate the c-statistic, and that we therefore used half of the 6865 participants and 312 events for estimating the standard error of the c-statistic.

2.2 Estimating the c-statistic from a reported D statistic

In some situations, validation studies report Royston and Sauerbrei's D statistic instead of the c-statistic. Because both measures are based on the standard deviation of the linear predictor (LP), they can be related to each other by assuming that the LP is Normally distributed.

2.2.1 Theoretical background

The following approximations were kindly provided by Prof. Ian White, and were derived from [25]:

1. Suppose that the LP is distributed according to $\text{LP} \sim \mathcal{N}(\mu_{\text{LP}}, \sigma^2_{\text{LP}})$.
2. Then the natural log hazard ratio (HR) for person j vs person i, assuming i and j are randomly selected, is distributed according to $\ln(\text{HR}) \sim \mathcal{N}(0, 2\,\sigma^2_{\text{LP}})$. So we can write

$$\ln(\text{HR}) = \sqrt{2}\,\sigma_{\text{LP}}\,z \tag{2}$$

   with $z \sim \mathcal{N}(0, 1)$.
3. Under a proportional hazards model, the probability that j's event occurs before i's event is given as $Pr(T_{\text{surv},j} < T_{\text{surv},i}) = \text{HR}/(\text{HR} + 1)$ where HR is for j compared to i.
4. This probability is the expit of ln(HR):

$$\text{HR}/(\text{HR} + 1) = \frac{\exp\left(\ln(\text{HR})\right)}{\exp\left(\ln(\text{HR})\right) + 1} \tag{3}$$

$$= \text{expit}\left(\ln(\text{HR})\right) \tag{4}$$

5. Hence the probability that a randomly selected pair are concordant is the average of $\text{expit}(\sqrt{2}\,\sigma_{\text{LP}}\,z)$ over $z \sim \mathcal{N}(0, 1)$ restricted to $z > 0$. Note that the restriction to $z > 0$ is necessary to obtain the probability of concordance (rather than discordance, for $z \leq 0$). This can also be rewritten as follows:

$$Pr(T_{\text{surv},j} < T_{\text{surv},i}) = \int_{-\infty}^{\infty} \text{expit}\left(\sqrt{2}\,\sigma_{\text{LP}}\,\text{abs}(z)\right)\phi(z)\,\partial z \tag{5}$$

$$= 2\int_{0}^{\infty} \text{expit}\left(\sqrt{2}\,\sigma_{\text{LP}}\,z\right)\phi(z)\,\partial z \tag{6}$$

   where $\phi(z)$ is the standard normal density function. This then yields the c-statistic as defined in equation (2) by White et al [25].

The relation between the standard deviation of the LP and Royston's D is as follows:

$$D = \sqrt{\frac{8}{\pi}}\sigma_{\text{LP}} \tag{7}$$

Hence, we can transform reported D values to estimates for the c-statistic by replacing σ_{LP} with $D\sqrt{\pi/8}$ in equation 6, yielding:

$$2\int_{0}^{\infty} \text{expit}\left(\sqrt{\frac{\pi}{4}}\,D\,z\right)\phi(z)\,\partial z \tag{8}$$

or simply

$$2\int_{0}^{\infty} \frac{\phi(z)}{1 + \exp\left(-0.5\sqrt{\pi}D\,z\right)}\,\partial z \tag{9}$$

2.2.2 Simulation Study

We conducted a simulation study to evaluate the accuracy of equation 6 in a logistic regression framework.

We take the approach of Austin and Steyerberg [26] and simulated a continuous explanatory variable $x_i \sim \mathcal{N}(0, \sigma_{\text{mc}})$ for each of $i = 1, \ldots, 1000$ subjects. The linear predictor is determined as follows:

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i}$$

where p_i denotes the probability of a binary condition occurring, $Pr(y_i = 1|x_{1i})$. For each subject, we then randomly generated a binary condition according to $y_i \sim \text{Bernoulli}(p_i)$. Afterwards, we fit a univariable logistic regression model (in which the binary condition was regressed on the continuous explanatory variable x_1) in the simulated dataset and estimated the c-statistic of the fitted model, which we refer to as the empirical c-statistic. We also determined the predicted c-statistic according to equation 6:

$$\hat{c}^{\dagger} = 2 \int_0^{\infty} \text{expit}\left(\sqrt{2}\,\hat{\sigma}_{\text{LP}}\, z\right) \phi(z)\, \partial z \qquad \text{(Approximation 1)}$$

and according to the equation previously described by Austin and Steyerberg [26]:

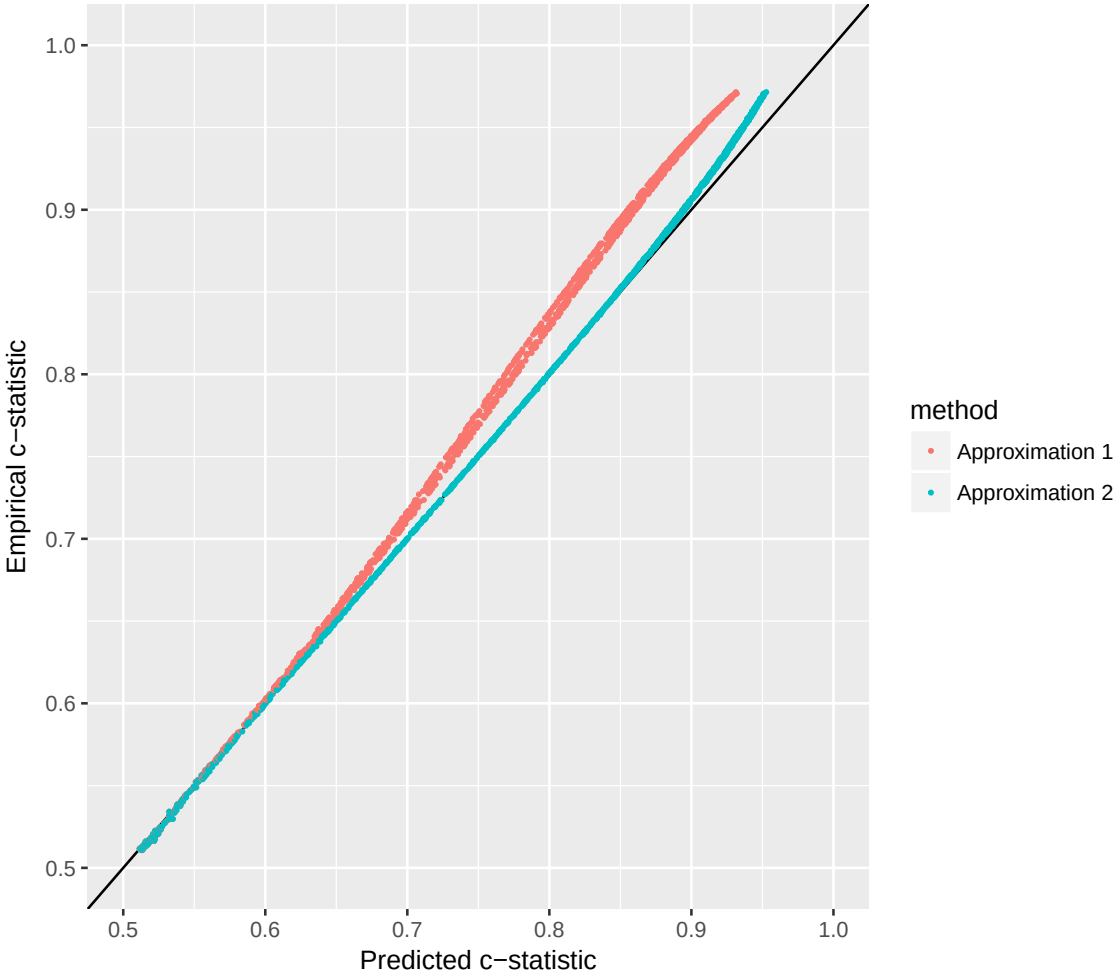
$$\hat{c}^{\ddagger} = \Phi\left(\frac{\hat{\mu}_{\text{A}} - \hat{\mu}_{\text{U}}}{\sqrt{\hat{\sigma}^2_{\text{A}} + \hat{\sigma}^2_{\text{U}}}}\right) \qquad \text{(Approximation 2)}$$

where $\hat{\mu}_{\text{A}}$ and $\hat{\mu}_{\text{U}}$ denote the estimated means, and where $\hat{\sigma}^2_{\text{A}}$ and $\hat{\sigma}^2_{\text{U}}$ denote the estimated variances of the linear predictor in the affected ($Y = 1$) and unaffected ($Y = 0$) populations, respectively. The estimated standard deviation of the linear predictor in the entire population is denoted as $\hat{\sigma}_{\text{LP}}$. Note that $\phi()$ and $\Phi()$ denote the density and, respectively, cumulative function of the standard normal distribution.

We performed a simulation study with a full factorial design in which the following factors were allowed to vary: β_0 (which influences the overall probability of the condition occurring), $\exp(\beta_1)$, and σ_{mc}. Hereby, we let β_0 to take on the values -2, -1, 0, 1, and 2; $\exp(\beta_1)$ to vary from 1 to 4 in increments of 0.2; and σ_{mc} to vary from 0.2 to 4 in increments of 0.2. In each of the 1,600 ($5 \times 16 \times 20$) different scenarios, we determined the mean of the empirical and predicted c-statistics in 100 simulated datasets.

The relationship between the predicted c-statistics and the empirical c-statistics across the scenarios is described in Figure S5. Approximation 1, which only requires information on the standard deviation of the linear predictor, provided accurate prediction of the c-statistic when the predicted c-statistic was less than 0.70. Conversely, when utilizing more detailed information on the distribution of the linear predictor, as done by Approximation 2, accurate prediction of the c-statistic were obtained when the predicted c-statistic was less than 0.90.

Figure S5: Comparison of empirical and predicted c-statistics



3 Data preparation of model calibration

3.1 Deriving event rates for prediction models with time-to-event outcomes

Let $S_{\mathrm{KM},t}$ denote the Kaplan-Meier estimate of the observed t-year cumulative survival probability, and $S_{\mathrm{E},t}$ denote the expected (cumulative) survival at t years. Note that $S_{\mathrm{E},t}$ can also be derived from the expected number of events and the total sample size. The total O:E ratio at t years is then given as:

$$(\mathrm{O{:}E})_t = \frac{1 - S_{\mathrm{KM},t}}{1 - S_{\mathrm{E},t}} \quad \Leftrightarrow \tag{10}$$

$$= \frac{1 - S_{\mathrm{KM},t}}{P_{\mathrm{E},t}} \tag{11}$$

where $P_{\mathrm{E,t}}$ represents the expected t-year cumulative event probabilities. If we treat the expected survival as a known quantity, we can approximate the error variance (Var) and the standard error (SE) of $(\mathrm{O{:}E})_t$ as follows:

$$\mathrm{Var}(\mathrm{O{:}E})_t = \mathrm{Var}\left(\frac{1 - S_{\mathrm{KM},t}}{1 - S_{\mathrm{E},t}}\right) \quad \Leftrightarrow \tag{12}$$

$$= \frac{1}{(1 - S_{\mathrm{E},t})^2}\,\mathrm{Var}\left(1 - S_{\mathrm{KM},t}\right) \quad \Leftrightarrow \tag{13}$$

$$= \frac{1}{(1 - S_{\mathrm{E},t})^2}\,\mathrm{Var}(S_{\mathrm{KM},t}) \quad \Leftrightarrow \tag{14}$$

$$= \frac{1}{(P_{\mathrm{E},t})^2}\,\mathrm{Var}(S_{\mathrm{KM},t}) \quad \Leftrightarrow \tag{15}$$

$$\mathrm{SE}(\mathrm{O{:}E})_t = \frac{1}{P_{\mathrm{E},t}}\,\mathrm{SE}(S_{\mathrm{KM},t}) \tag{16}$$

such that:

$$\mathrm{Var}(\ln(\mathrm{O{:}E})_t) = \mathrm{Var}\left(\ln\left(\frac{1 - S_{\mathrm{KM},t}}{1 - S_{\mathrm{E},t}}\right)\right) \quad \Leftrightarrow \tag{17}$$

$$= \mathrm{Var}\left(\ln\left(1 - S_{\mathrm{KM},t}\right) - \ln(1 - S_{\mathrm{E},t})\right) \quad \Leftrightarrow \tag{18}$$

$$= \mathrm{Var}\left(\ln\left(1 - S_{\mathrm{KM},t}\right)\right) + \mathrm{Var}(\ln(1 - S_{\mathrm{E},t})) \quad \Leftrightarrow \tag{19}$$

$$= \mathrm{Var}\left(\ln\left(1 - S_{\mathrm{KM},t}\right)\right) \quad \Leftrightarrow \tag{20}$$

$$\approx \left(\frac{\partial\left[\ln\left(1 - S_{\mathrm{KM},t}\right)\right]}{\partial\left[S_{\mathrm{KM},l}\right]}\right)^2 \mathrm{Var}(S_{\mathrm{KM},l}) \quad \Leftrightarrow \tag{21}$$

$$\approx \frac{1}{(1 - S_{\mathrm{KM},l})^2}\mathrm{Var}(S_{\mathrm{KM},l}) \quad \Leftrightarrow \tag{22}$$

$$\mathrm{SE}(\ln(\mathrm{O{:}E})_t) \approx \frac{1}{1 - S_{\mathrm{KM},l}}\mathrm{SE}(S_{\mathrm{KM},l}) \tag{23}$$

If $\mathrm{SE}(S_{\mathrm{KM},t})$ is not reported, we can use the standard error for a population proportion:

$$\mathrm{SE}(S_{\mathrm{KM},t}) \approx \sqrt{\frac{S_{\mathrm{KM},t}(1 - S_{\mathrm{KM},t})}{N_t}} \tag{24}$$

with N_t the number of participants with complete outcome information (i.e. whether an event occurred or not) after t years of follow-up.

3.2 Extrapolation of event rates

In some situations, we do not have survival estimates at specific time points of interest. For instance, we may have survival estimates for l (e.g. 5 years) instead of t (e.g. 10 years). Below, we describe how to extrapolate Kaplan-Meier event rates across different time points. If we assume that events occur according to a Poisson distribution, we have for a given l:

$$\begin{aligned} S_{\mathrm{KM},t} &= \mathrm{Poisson}(0;\lambda) && \Leftrightarrow && (25)\\ &= \mathrm{Poisson}(0;\phi t) && \Leftrightarrow && (26)\\ &= \exp(-\phi t) && \Leftrightarrow && (27)\\ &= \exp\left(t \ln(S_{\mathrm{KM},l})/l\right) && && (28) \end{aligned}$$

where λ represents the shape parameter which indicates the average number of events in the given time interval t. The parameter ϕ then represents the "normalized rate" in time units of t (e.g., per year, month or second).

For instance, the observed 5-year cumulative survival probability $S_{\mathrm{KM},5}$ can be extrapolated to the 10-year cumulative survival probability using:

$$S_{\mathrm{KM},10} = \exp\left(\frac{10\ \ln(S_{\mathrm{KM},5})}{5}\right) \tag{29}$$

We can approximate the standard error of extrapolated event rates using the Delta method:

$$\begin{aligned} \mathrm{SE}(S_{\mathrm{KM},t}) &\approx \sqrt{\left(\frac{\partial\left[\exp\left(t \ln(S_{\mathrm{KM},l})/l\right)\right]}{\partial\left[S_{\mathrm{KM},l}\right]}\right)^2}\mathrm{SE}(S_{\mathrm{KM},l}) && \Leftrightarrow && (30)\\ &\approx \sqrt{\left(\frac{t\exp\left(t \ln(S_{\mathrm{KM},l})/l\right)}{l\ S_{\mathrm{KM},l}}\right)^2}\mathrm{SE}(S_{\mathrm{KM},l}) && \Leftrightarrow && (31)\\ &\approx \frac{t\exp\left(t \ln(S_{\mathrm{KM},l})/l\right)}{l\ S_{\mathrm{KM},l}}\ \mathrm{SE}(S_{\mathrm{KM},l}) && && (32) \end{aligned}$$

The total O:E ratio at t years can then be calculated as:

$$\begin{aligned} (\mathrm{O{:}E})_t &= \frac{1-S_{\mathrm{KM},t}}{1-S_{\mathrm{E},t}} && \Leftrightarrow && (33)\\ &= \frac{1-\exp(t \ln(S_{\mathrm{KM},l})/l)}{1-\exp(t \ln(S_{\mathrm{E},l})/l)} && && (34) \end{aligned}$$

The error variance of $(\mathrm{O{:}E})_t$ can be approximated as follows:

$$\begin{aligned} \mathrm{Var}(\mathrm{O{:}E})_t &= \mathrm{Var}\left(\frac{1-\exp(t \ln(S_{\mathrm{KM},l})/l)}{1-\exp(t \ln(S_{\mathrm{E},l})/l)}\right) && \Leftrightarrow && (35)\\ &= \left(\frac{1}{1-\exp(t \ln(S_{\mathrm{E},l})/l)}\right)^2 \mathrm{Var}\left(1-\exp(t \ln(S_{\mathrm{KM},l})/l)\right) && \Leftrightarrow && (36)\\ &\approx \frac{1}{(1-\exp(t \ln(S_{\mathrm{E},l})/l))^2}\ \left(\frac{\partial\left[(1-\exp(t \ln(S_{\mathrm{KM},l})/l))\right]}{\partial\left[S_{\mathrm{KM},l}\right]}\right)^2 \mathrm{Var}(S_{\mathrm{KM},l}) && \Leftrightarrow && (37)\\ &\approx \frac{1}{(1-\exp(t \ln(S_{\mathrm{E},l})/l))^2}\ \left(-\frac{t\exp(t \ln(S_{\mathrm{KM},l})/l)}{lS_{\mathrm{KM},l}}\right)^2 \mathrm{Var}(S_{\mathrm{KM},l}) && \Leftrightarrow && (38)\\ &\approx \frac{t^2\exp(2t \ln(S_{\mathrm{KM},l})/l)}{l^2(S_{\mathrm{KM},l})^2(1-\exp(t \ln(S_{\mathrm{E},l})/l))^2}\mathrm{Var}(S_{\mathrm{KM},l}) && \Leftrightarrow && (39)\\ \mathrm{SE}(\mathrm{O{:}E})_t &\approx \frac{t\exp(t \ln(S_{\mathrm{KM},l})/l)}{lS_{\mathrm{KM},l}(1-\exp(t \ln(S_{\mathrm{E},l})/l))}\mathrm{SE}(S_{\mathrm{KM},l}) && && (40) \end{aligned}$$

which can further be simplified to:

$$\begin{aligned}
\mathrm{Var}(\mathrm{O{:}E})_t &\approx \frac{t^2 \exp(2t \ln(S_{\mathrm{KM},l})/l)}{l^2 (S_{\mathrm{KM},l})^2 (1 - \exp(t \ln(S_{\mathrm{E},l})/l))^2} \mathrm{Var}(S_{\mathrm{KM},l}) && \Leftrightarrow && (41)\\
&\approx \frac{t^2 \exp(2t \ln(S_{\mathrm{KM},l})/l)}{l^2 (S_{\mathrm{KM},l})^2 (1 - \exp(t \ln(S_{\mathrm{E},l})/l))^2} \left(\frac{S_{\mathrm{KM},l}(1 - S_{\mathrm{KM},l})}{N_l} \right) && \Leftrightarrow && (42)\\
&\approx \frac{t^2 \exp(2t \ln(S_{\mathrm{KM},l})/l)(1 - S_{\mathrm{KM},l})}{l^2 N_l S_{\mathrm{KM},l} (1 - \exp(t \ln(S_{\mathrm{E},l})/l))^2} && \Leftrightarrow && (43)\\
\mathrm{SE}(\mathrm{O{:}E})_t &\approx \frac{t \sqrt{(1 - S_{\mathrm{KM},l})} \exp(t \log(S_{\mathrm{KM},l})/l)}{l \sqrt{N_l S_{\mathrm{KM},l}} \, (1 - \exp(t \log(S_{\mathrm{E},l})/l))} && && (44)
\end{aligned}$$

For the natural logarithm of the total O:E ratio we have:

$$\begin{aligned}
\ln(\mathrm{O{:}E})_t &= \ln \left(\frac{1 - \exp(t \ln(S_{\mathrm{KM},l})/l)}{1 - \exp(t \ln(S_{\mathrm{E},l})/l)} \right) && \Leftrightarrow && (45)\\
&= \ln \left(1 - \exp(t \ln(S_{\mathrm{KM},l})/l)\right) - \ln \left(1 - \exp(t \ln(S_{\mathrm{E},l})/l)\right) && && (46)
\end{aligned}$$

with

$$\begin{aligned}
\mathrm{Var}(\ln(\mathrm{O{:}E})_t) &= \mathrm{Var} \left(\ln \left(1 - \exp(t \ln(S_{\mathrm{KM},l})/l)\right) - \ln \left(1 - \exp(t \ln(S_{\mathrm{E},l})/l)\right) \right) && \Leftrightarrow && (47)\\
&= \mathrm{Var} \left(\ln \left(1 - \exp(t \ln(S_{\mathrm{KM},l})/l)\right) \right) && \Leftrightarrow && (48)\\
&\approx \left(\frac{\partial \left[\ln \left(1 - \exp(t \ln(S_{\mathrm{KM},l})/l)\right) \right]}{\partial \left[S_{\mathrm{KM},l} \right]} \right)^2 \mathrm{Var}(S_{\mathrm{KM},l}) && \Leftrightarrow && (49)\\
&\approx \frac{t^2 \exp\left(2t \ln(S_{\mathrm{KM},l})/l\right)}{l^2 (S_{\mathrm{KM},l})^2 (1 - \exp\left(t \ln(S_{\mathrm{KM},l})/l\right))^2} \mathrm{Var}(S_{\mathrm{KM},l}) && \Leftrightarrow && (50)\\
\mathrm{SE}(\ln(\mathrm{O{:}E})_t) &\approx \left(\frac{t}{l S_{\mathrm{KM},l}} \right) \left(\frac{\exp\left(t \ln(S_{\mathrm{KM},l})/l\right)}{1 - \exp\left(t \ln(S_{\mathrm{KM},l})/l\right)} \right) \mathrm{SE}(S_{\mathrm{KM},l}) && && (51)
\end{aligned}$$

Note that the SE of the observed cumulative l-year survival, $\mathrm{SE}(S_{\mathrm{KM},l})$, is equal to the SE of the observed cumulative l-year risk, $\mathrm{SE}(1 - S_{\mathrm{KM},l})$. If both quantities are unavailable, aforementioned equations can again be simplified:

$$\begin{aligned}
\mathrm{Var}(\ln(\mathrm{O{:}E})_t) &\approx \frac{t^2 \exp\left(2t \ln(S_{\mathrm{KM},l})/l\right)}{l^2 (S_{\mathrm{KM},l})^2 (1 - \exp\left(t \ln(S_{\mathrm{KM},l})/l\right))^2} \mathrm{Var}(S_{\mathrm{KM},l}) && \Leftrightarrow && (52)\\
&\approx \frac{t^2 \exp\left(2t \ln(S_{\mathrm{KM},l})/l\right)}{l^2 (S_{\mathrm{KM},l})^2 (1 - \exp\left(t \ln(S_{\mathrm{KM},l})/l\right))^2} \left(\frac{S_{\mathrm{KM},l}(1 - S_{\mathrm{KM},l})}{N_l} \right) && \Leftrightarrow && (53)\\
&\approx \frac{t^2 (1 - S_{\mathrm{KM},l}) \exp\left(2t \ln(S_{\mathrm{KM},l})/l\right)}{l^2 N_l S_{\mathrm{KM},l} (1 - \exp\left(t \ln(S_{\mathrm{KM},l})/l\right))^2} && \Leftrightarrow && (54)\\
\mathrm{SE}(\ln(\mathrm{O{:}E})_t) &\approx \frac{t \sqrt{(1 - S_{\mathrm{KM},l})} \exp(t \log(S_{\mathrm{KM},l})/l)}{l \sqrt{N_l S_{\mathrm{KM},l}} \, (1 - \exp(t \log(S_{\mathrm{KM},l})/l))} && && (55)
\end{aligned}$$

3.3 Examples

For all examples, we consider validation of Framingham Wilson in a male population.

Example 1: DeFilippis 2015

In the first example [9], we consider the situation where $O_{t=10} = 164$, $E_{t=10} = 251.1$ and $P_{\mathrm{O},t=10} = 0.084$ have been reported (see Table 3 in the publication). Hence, we can calculate $(\mathrm{O{:}E})_{t=10}$ fairly straightforward:

$$(\mathrm{O{:}E})_{t=10} = 164/251.1 = 0.65 \tag{56}$$

such that

$$\ln(\mathrm{O{:}E})_{t=10} = \ln(164) - \ln(251.1) = -0.426 \tag{57}$$

The standard error of $\ln(\mathrm{O{:}E})_{t=10}$ is given as [27]:

$$\mathrm{SE}(\ln(\mathrm{O{:}E})_{t=10}) = \sqrt{(1 - 0.084)/164} = 0.075 \tag{58}$$

Alternatively, if $P_{\mathrm{O},t=10}$ was not reported, the standard error could be approximated as [27]:

$$\mathrm{SE}(\ln(\mathrm{O{:}E})_{t=10}) = \sqrt{1/164} = 0.078 \tag{59}$$

Example 2: Buitrago 2011

In the validation by [6], the observed and predicted 10-year risk estimates were reported (Table 2), as well as the total number of observed events (Table 1). The resulting summary data is given as $P_{\mathrm{O},t=10} = 0.109$, $P_{\mathrm{E},t=10} = 0.169$ and $O_{t=10} = 22$. Hence, we have:

$$(\mathrm{O{:}E})_{t=10} = 0.109/0.169 = 0.645 \tag{60}$$

such that

$$\ln(\mathrm{O{:}E})_{t=10} = \ln(0.109) - \ln(0.169) = -0.439 \tag{61}$$

with

$$\mathrm{SE}(\ln(\mathrm{O{:}E})_{t=10}) = \sqrt{(1 - 0.109)/22} = 0.201 \tag{62}$$

Example 3: Rodondi 2012

In the third example [17], event rates are only reported for a follow-up of 7.5 rather than 10 years. In particular, we have have $P_{\mathrm{O},t=7.5} = 0.2071$ and $P_{\mathrm{E},t=7.5} = 0.1903$ (Table S1). We can extrapolate these event rates as follows:

$$S_{\mathrm{KM},t=10} = \exp\left(10 \ln(1 - 0.2071)/7.5\right) = 0.734 \tag{63}$$
$$S_{\mathrm{E},t=10} = \exp\left(10 \ln(1 - 0.1903)/7.5\right) = 0.755 \tag{64}$$

such that:

$$(\mathrm{O{:}E})_{t=10} = (1 - 0.734)/(1 - 0.755) = 1.08 \tag{65}$$
$$(\mathrm{O{:}E})_{t=7.5} = 0.2071/0.1903 = 1.09 \tag{66}$$

and

$$\ln(\mathrm{O{:}E})_{t=10} = \ln(1 - 0.734) - \ln(1 - 0.755) = 0.082 \tag{67}$$

The standard error is given as:

$$\mathrm{SE}(\ln(\mathrm{O{:}E})_{t=10}) = \left(\frac{10}{7.5\,(1-0.2071)}\right)\left(\frac{\exp\left(10\ln(1-0.2071)/7.5\right)}{1-\exp\left(10\,\ln(1-0.2071)/7.5\right)}\right)\mathrm{SE}(S_{\mathrm{KM},l}) \tag{68}$$

$$= 1.681591 \times 2.757691 \times \mathrm{SE}(S_{\mathrm{KM},l}) \tag{69}$$

$$\approx 1.681591 \times 2.757691 \times \sqrt{\frac{(1-0.2071)\times 0.2071}{N_{t=7.5}}} \tag{70}$$

Although $N_{t=7.5}$ is not directly reported, we here assume that $N_{t=7.5} = N = 981$. This assumption seems reasonable because the number of observed events over the entire follow-up ($O = 205$) is similar to what can be estimated from $P_{\mathrm{O},t=7.5}$ assuming that there is no drop-out ($0.2071 \times 981 = 203$). We can therefore approximate $\mathrm{SE}(\ln(\mathrm{O{:}E})_{t=10})$ as follows:

$$\mathrm{SE}(\ln(\mathrm{O{:}E})_{t=10}) \approx 1.681591 \times 2.757691 \times \sqrt{\frac{(1-0.2071)\times 0.2071}{981}} \tag{71}$$

$$\approx 1.681591 \times 2.757691 \times 0.01293793 \tag{72}$$

$$\approx 0.06 \tag{73}$$

Note that estimate is almost identical to the binomial approximation ignoring extrapolation of $P_{\mathrm{O},t}$ (eq. 27 from [27]):

$$\mathrm{SE}(\ln(\mathrm{O{:}E})_{t=10}) \approx \sqrt{\frac{1-(1-0.734)}{203}} \tag{74}$$

$$\approx 0.06 \tag{75}$$

Finally, note that in studies with substantial drop-out of participants, it is still possible to estimate N_l by assuming that patients are censored at a constant rate [28, 29].

Example 4: Empana 2003

In the fourth example by Empana [10], 5-year event rates are depicted graphically for tenths of predicted risk (see Figure S6). The total sample size was of the validation study was 2399. If we extract the different calibration points (see Table S5), we can estimate the total O:E ratio as follows:

$$P_{\mathrm{O},t=5} = \frac{2+5+11+9+6+10+15+15+26+21}{2399} = 0.05002084 \tag{76}$$

$$P_{\mathrm{E},t=5} = \frac{6+9+11+12+14+16+18+21+24+35}{2399} = 0.0691955 \tag{77}$$

such that $(\mathrm{O{:}E})_{t=5} = 120/166 = 0.72$, which is very close to the ratio reported in the original validation study (as indicated in Figure 1 of [10], this ratio was $1/1.34 = 0.75$). We can extrapolate this calibration statistic to a 10-year follow-up period as follows:

$$(\mathrm{O{:}E})_{t=10} = \frac{1-\exp\left(10\,\ln(1-0.05002084)/5\right)}{1-\exp\left(10\,\ln(1-0.0691955)/5\right)} = \frac{0.0975}{0.1336} = 0.73 \tag{78}$$

such that $\ln(\mathrm{O{:}E})_{t=10} = -0.315$. The standard error can be approximated using equation (27) from the Appendix of [27], such that:

$$\mathrm{SE}(\ln(\mathrm{O{:}E})_{t=10}) \approx \sqrt{\frac{1-0.0975}{120}} \approx 0.087 \tag{79}$$

or, more accurately (assuming that $N_{t=5} = 2399$):

$$\mathrm{SE}(\ln(\mathrm{O{:}E})_{t=10}) = \left(\frac{10}{5\,(1-0.0975)}\right)\left(\frac{\exp\left(10\ln(1-0.0975)/5\right)}{1-\exp\left(10\,\ln(1-0.0975)/5\right)}\right)\sqrt{\frac{(1-0.0975)\,0.0975}{2399}} \tag{80}$$

$$= 0.087 \tag{81}$$

As an illustration, estimates for $\mathrm{SE}(\ln(\mathrm{O{:}E})_{t=10})$ are plotted against values for $N_{t=5}$ in Figure S7. For instance, if $N_{t=5}$ would be 2000 rather than 2399 (e.g. due to drop-out of study participants), the SE increases from 0.087 to 0.095.

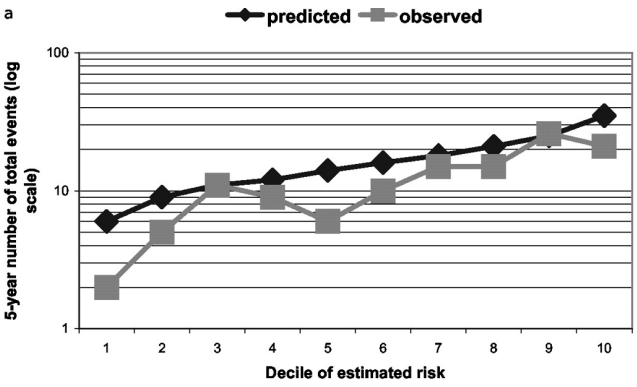
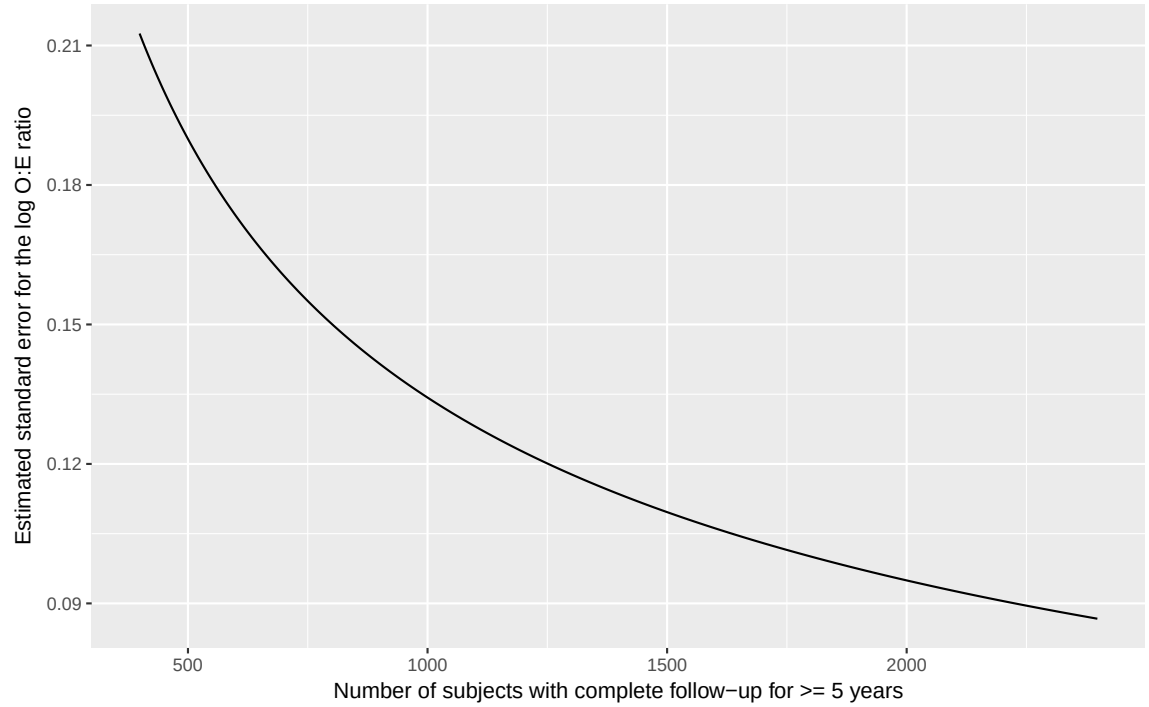


Figure S6: Calibration of the Framingham risk function in Belfast for tenths of predicted risk, as obtained from Empana 2003.

	Total	**Observed**		**Predicted**	
	N	N	%	N	%
1	240	2	(0.83 %)	6	(2.50 %)
2	240	5	(2.08 %)	9	(3.75 %)
3	240	11	(4.59 %)	11	(4.59 %)
4	240	9	(3.75 %)	12	(5.00 %)
5	240	6	(2.50 %)	14	(5.84 %)
6	240	10	(4.17 %)	16	(6.67 %)
7	240	15	(6.25 %)	18	(7.50 %)
8	240	15	(6.25 %)	21	(8.75 %)
9	240	26	(10.84 %)	24	(10.00 %)
10	239	21	(8.75 %)	35	(14.59 %)

Table S5: Extracted event rates for the validation of Framingham Wilson in Empana 2003.

Figure S7: Estimates for $\mathrm{SE}(\ln(\mathrm{O{:}E})_{t=10})$ in example 4.



4 Bayesian Meta-analysis

4.1 Empirical estimates of between-study heterogeneity

Table S6: Empirical estimates of between-study heterogeneity for meta-analysis of the c-statistic

Study	**Meta-analysis scale**			
	Original		Logit	
	Mean	BSSD	Mean	BSSD
Kengne 2004 † [30]	0.80	0	1.39	0
Thompson 2014 [31]	0.60	0.001	0.41	0.00
Ford 2010 [32]	0.64	0.018	0.57	0.00
Thompson 2014 [31]	0.62	0.022	0.48	0.09
Snell 2015 † [33]	0.71	0.019	0.80	0.11
Ford 2010 [32]	0.76	0.011	1.08	0.11
Debray 2017 [27]	0.79	0.029	1.31	0.15
Gagne 2015 [34]	0.83	0.027	1.57	0.18
Kengne 2004 † [30]	0.76	0.037	1.15	0.21
Kengne 2004 † [30]	0.77	0.045	1.21	0.25
Kengne 2004 † [30]	0.79	0.044	1.31	0.26
Guida 2014 [35]	0.79	0.045	1.32	0.26
Kengne 2004 † [30]	0.79	0.046	1.31	0.28
Kengne 2004 † [30]	0.79	0.046	1.31	0.28
Meads 2011 † [36]	0.63	0.068	0.53	0.29
Chalmers 2011 [37]	0.78	0.050	1.27	0.29
Shen 2016 [38]	0.69	0.068	0.79	0.30
Chalmers 2011 [37]	0.66	0.066	0.68	0.31
Shen 2016 [38]	0.78	0.057	1.23	0.32
Chalmers 2011 [37]	0.79	0.052	1.34	0.33
Chalmers 2011 [37]	0.67	0.075	0.71	0.33
Marques 2015 [39]	0.79	0.062	1.34	0.36
Ford 2010 [32]	0.65	0.095	0.64	0.44
Gagne 2015 [34]	0.83	0.063	1.68	0.45
van Klaveren 2014 † [40]	0.77	0.085	1.21	0.48
Marques 2015 [39]	0.74	0.090	1.08	0.49

Mean = random effects summary estimate; BSSD = between-study standard deviation

† Estimates for the BSSD on the transformed scale were obtained by applying the Delta method: $\text{Var}(\text{logit}(c)) = \text{Var}(c)/(c\,(1-c)^2)$, where we assumed that $\text{Var}(c)$ is the between-study variance of the c-statistic.

Table S7: Empirical estimates of between-study heterogeneity for meta-analysis of the total O:E ratio.

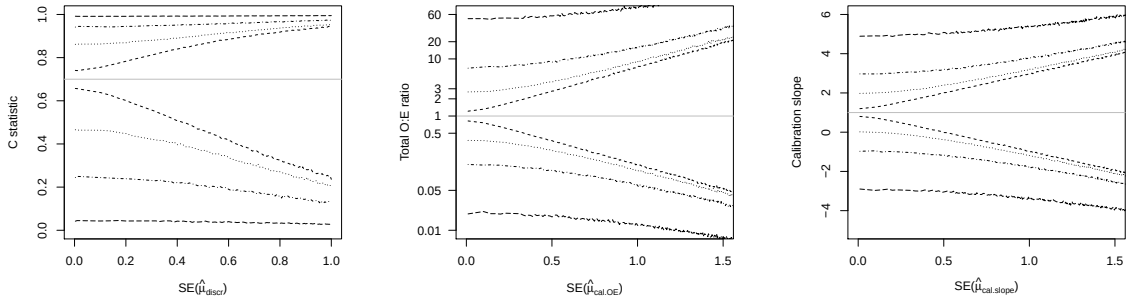
Review		**Meta-analysis scale**			
		Original		Log	
	N	Mean	BSSD	Mean	BSSD
Ohle 2011 [41]	5	0.12	0	-1.61	0
Zhu 2015 [42]	5	0.75	0	-0.11	0
Ohle 2011 [41]	6	0.62	0.36	-0.04	0
Ohle 2011 [41]	7	1.01	0.05	0.01	0.05
Meads 2011 [36]	10	1.05	0.09	0.06	0.09
Ohle 2011 [41]	10	0.90	0.14	-0.12	0.16
Ohle 2011 [41]	4	0.93	0.15	-0.05	0.18
Ohle 2011 [41]	8	0.92	0.20	-0.09	0.20
Zhu 2015 [42]	6	1.12	0.27	0.14	0.20
Meads 2011 [36]	3	0.90	0.24	-0.12	0.26
Ohle 2011 [41]	4	0.53	0.17	-0.59	0.31
Zhu 2015 [42]	6	1.40	0.54	0.32	0.37
Guida 2014 [35]	23	0.99	0.00	0.10	0.44
Debray 2017 [27]	19	0.57	0.25	-0.64	0.47
Ohle 2011 [41]	7	0.54	0.20	-0.49	0.55
Ohle 2011 [41]	5	0.47	0.62	-1.15	1.39

N = number of included studies; Mean = random effects summary estimate; BSSD = between-study standard deviation

Summary estimates were obtained using random effects meta-analysis using restricted maximum likelihood estimation.

4.2 Influence of between-study heterogeneity

Figure S8: Prediction intervals for varying degrees of between-study heterogeneity



Prediction intervals were generated by adopting a Bayesian estimation framework where we assumed fixed values for the presence of between-study heterogeneity, according to $\tau = 0.1$ (dash), $\tau = 0.5$ (dot), $\tau = 1$ (dot-dash) and $\tau = 2$ (long dash).

4.3 Construction of prior distributions

Based on the empirical data presented in Section 4.1, we consider the construction of weakly informative prior distributions for the between-study standard deviation of the logit c-statistic(τ_{discr}) and of the log O:E ratio ($\tau_{\text{cal.OE}}$). Results in Section 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, $\tau_{\text{cal.OE}} > 2$. Hence, we propose the following priors adopting a Uniform distribution:

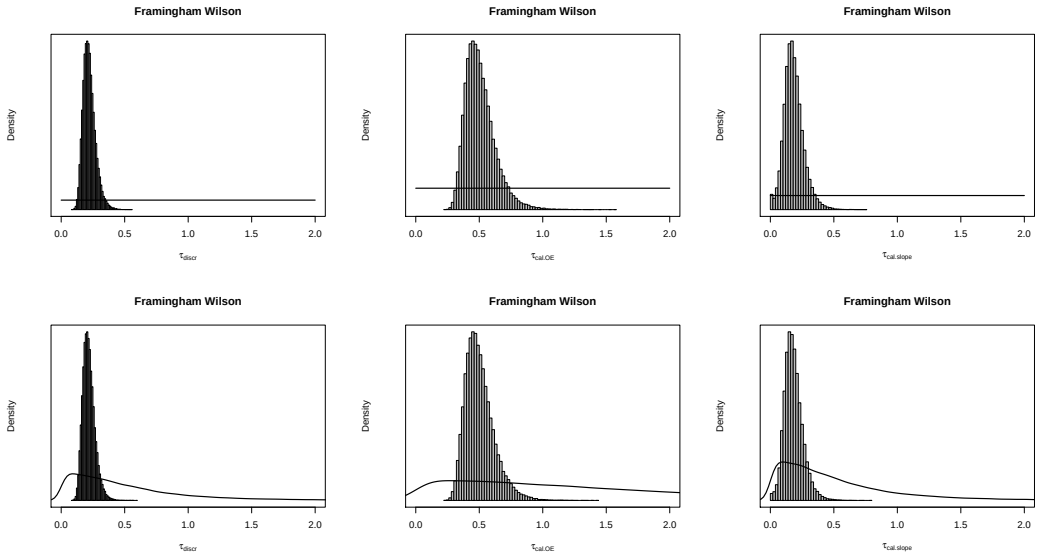
- $\tau_{\text{discr}} \sim \text{Unif}(0, 2)$
- $\tau_{\text{cal.OE}} \sim \text{Unif}(0, 2)$
- $\tau_{\text{cal.slope}} \sim \text{Unif}(0, 2)$

Because the uniform distribution tends to unduly favor presence of heterogeneity in discrimination and calibration estimates across studies [43, 44], we also consider a half Student-t distribution with location m, scale σ and ν degrees of freedom:

- $\tau_{\text{discr}} \sim \text{Student-}t(m = 0, \sigma = 0.5, \nu = 3)\, T[0, 10]$
- $\tau_{\text{cal.OE}} \sim \text{Student-}t(m = 0, \sigma = 1.5, \nu = 3)\, T[0, 10]$
- $\tau_{\text{cal.slope}} \sim \text{Student-}t(m = 0, \sigma = 0.5, \nu = 3)\, T[0, 10]$

The resulting priors are depicted in the figure below for meta-analysis of the Framingham Risk Score. Hereby, we assumed within-study Normality for the logit c-statistic and, respectively, log O:E ratio.

Figure S9: Histograms of posterior simulations of the between-study standard deviation in the empirical examples.



The histogram represents the posterior on τ, whereas the solid line indicates the prior density. In the top figures, a Uniform prior is used. In the bottom figures, a truncated Student-t prior is used.

4.4 Meta-analysis models

Below, we describe the meta-analysis models for obtaining summary estimates of the c-statistic, the total O:E ratio, and the calibration slope. To facilitate their implementation, we integrated all methods in the R package `metamisc`. This package currently supports restoring of missing information (e.g. standard error of the c-statistic), necessary data transformations, and frequentist and Bayesian meta-analysis of the c-statistic and total O:E ratio. Methods for summarizing the calibration slope will be implemented in the future, however, corresponding source code is already available from the supporting information.

To use the package `metamisc`, apply the following commands in R:

```
install.packages("metamisc")
library("metamisc")
packageVersion("metamisc") # Verify package version
```

The examples below were performed using `metamisc` version 0.1.8.

4.4.1 Meta-analysis of the c-statistic

The marginal model for a frequentist meta-analysis of the c-statistic is given as follows:

$$\begin{aligned} \text{logit}(c_i) &\sim \mathcal{N}\left(\theta_i, \text{Var}\left(\text{logit}(c_i)\right)\right) \\ \theta_i &\sim \mathcal{N}\left(\mu_{\text{discr}}, \tau^2_{\text{discr}}\right) \end{aligned} \qquad \text{(Model 1)}$$

The corresponding R code to summarize the discrimitive performance of EuroSCORE II is given below.

```
data(EuroSCORE)
fit <- with(EuroSCORE, valmeta(cstat=c.index, cstat.se=se.c.index,
               cstat.95CI=cbind(c.index.95CIl,c.index.95CIu),
               N=n, O=n.events, slab=Study))
```

By default, `metamisc` adopts restricted maximum likelihood estimation and uses the Sidik-Jonkman Hartung-Knapp method for constructing confidence intervals. This results in the following estimates for the meta-analysis of EuroSCORE II:

```
> fit
Model results for the c-statistic:

 estimate     95CIl     95CIu     95PIl     95PIu
0.7888603 0.7648784 0.8110005 0.6795982 0.8680942

Number of studies included:  23
Note: For 4 validation(s), the standard error of the concordance statistic
was estimated using method 'Newcombe.4'.
```

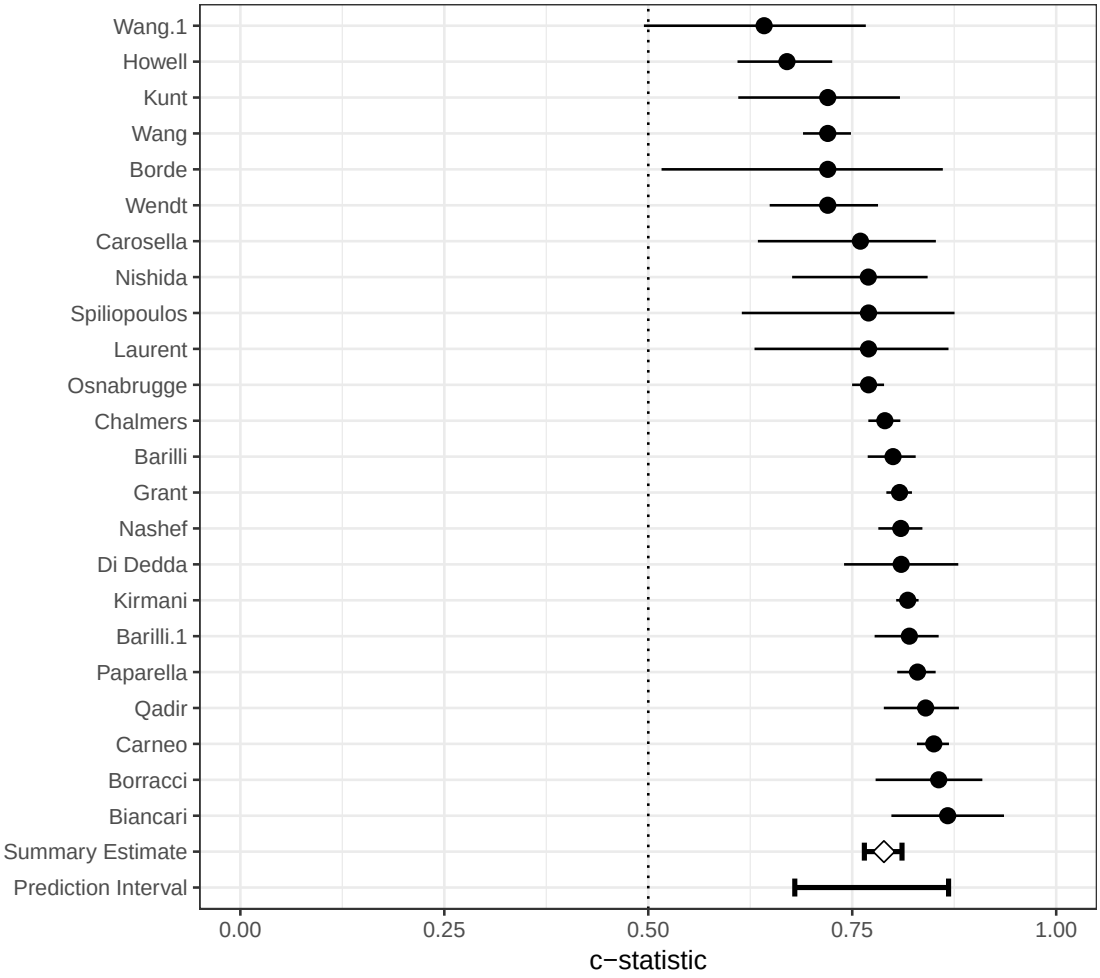
When some studies did not provide a standard error for the (logit) c-statistic, it is often helpful to inspect how this quantity was estimated using the `valmeta` command. This information, together with the data used for meta-analysis, is available from `fit$data`.

```
> head(fit$data)
     theta   theta.se theta.CIl theta.CIu theta.blup     theta.se.source
1 1.446765 0.08964514  1.277360  1.628762   1.433523 Confidence Interval
2 1.874690 0.33390703  1.373841  2.682732   1.532874 Confidence Interval
3 1.450010 0.24144872  1.045969  1.992430   1.390088 Confidence Interval
4 1.324925 0.06027728  1.206784  1.443067   1.324587      Standard Error
5 1.437067 0.05156766  1.335996  1.538137   1.432715      Standard Error
6 1.734601 0.07843137  1.580878  1.888324   1.700983      Standard Error
```

Finally, it is possible to generate a forest plot (see Figure S10).

```
plot(fit)
```

Figure S10: Forest plot of the discrimination performance of EuroSCORE II



We can consider a Bayesian meta-analysis to account for uncertainty in the estimation of τ_{discr}. By default, it is assumed that $\mu_{\text{discr}} \sim \mathcal{N}(0, 10^6)$ and $\tau_{\text{discr}} \sim \text{Unif}(0, 2)$. This model can be implemented as follows:

```
fit <- with(EuroSCORE, valmeta(cstat=c.index, cstat.se=se.c.index,
                cstat.95CI=cbind(c.index.95CIl,c.index.95CIu),
                N=n, O=n.events, slab=Study, method="BAYES"))
```

Note that meta-analysis results are very similar to the ones obtained with REML estimation:

```
> fit
Summary c-statistic with 95% credibility and  95% prediction interval:

 estimate       CIl       CIu       PIl       PIu
0.7884944 0.7639814 0.8115765 0.6826630 0.8809185

Penalized expected deviance:  97.49

Number of studies included:  23
Note: For 4 validation(s), the standard error of the concordance statistic
was estimated using method 'Newcombe.4'.
```

The fitted `JAGS` model can be retrieved from `fit$runjags`. This object can subsequently be used to investigate convergence in more detail, or to print various characteristics:

```
> print(fit$runjags$model)

JAGS model syntax:
```

```
1   |  model{
2   |  for (i in 1:Nstudies)
3   |  {
4   |  theta[i] ~ dnorm(alpha[i], wsprec[i])
5   |  alpha[i] ~ dnorm(mu.tobs, bsprec)
6   |  wsprec[i] <- 1/(theta.var[i])
7   |  }
8   |  bsprec <- 1/(bsTau*bsTau)
9   |  bsTau ~ dunif(0,2)
10  |  mu.tobs ~ dnorm(0,1e-06)
11  |  mu.obs <- 1/(1+exp(-mu.tobs))
12  |  pred.obs <- 1/(1+exp(-pred.tobs))
13  |  pred.tobs ~ dnorm(mu.tobs, bsprec)
14  |  }

> fit$runjags$data

JAGS data:

1  |  "theta" <- c(1.44676457855093, 1.87468984855879, 1.450010175506,
 1.3249254147436, 1.43706668649331, 1.73460105538811, 0.944461608840851,
 1.50285564952595, 0.708185057924486, 0.944461608840851, 0.944461608840851,
 1.65822807660353, 1.20831120592453, 0.944461608840851, 1.20831120592453,
 0.584055317289261, 1.20661802171363, 1.38629436111989, 1.51634748936809,
 1.58562726374038, 1.15267950993839, 1.78245707656574, 1.20831120592453)
2  |  "theta.var" <- c(0.00803625185405368, 0.111493902485701, 0.0582974867155871,
 0.00363334993774255, 0.00265922322430263, 0.00615148019992311, 0.0639969529478458,
 0.00221078595421536, 0.0180780965840686, 0.00553606859410431, 0.201277475968239,
 0.0302664009358649, 0.14312400583974, 0.0284430901990426, 0.118637653314697,
 0.0954290472715787, 0.0574270742259427, 0.0087890625, 0.0183606172105081,
 0.00723283344743463, 0.0942597722376116, 0.071672964220553, 0.00318832715169838)
3  |  "Nstudies" <- 23
4  |
```

We proposed $\tau_{\text{discr}} \sim$ Student-$t(0, 0.5^2, 3)\, T[0, 10]$ as an alternative prior for the between-study standard deviation. This prior can be implemented as follows:

```
pars.model <- list(hp.tau.dist="dhalft",
                   hp.tau.sigma=0.5,
                   hp.tau.min=0,
                   hp.tau.max=10,
                   hp.tau.df=3)
with(EuroSCORE, valmeta(cstat=c.index, cstat.se=se.c.index, N=n, O=n.events,
                        cstat.95CI=cbind(c.index.95CIl, c.index.95CIu),
                        slab=Study, method="BAYES", pars=pars.model))
```

4.4.2 Meta-analysis of the total O:E ratio

For meta-analysis of the total O:E ratio we have:

$$\begin{aligned} \ln(\text{O:E})_i &\sim \mathcal{N}\left(\zeta_i, \text{Var}\left(\ln(\text{O:E})_i\right)\right) \\ \zeta_i &\sim \mathcal{N}\left(\mu_{\text{cal.OE}}, \tau^2_{\text{cal.OE}}\right) \end{aligned} \tag{Model 2}$$

with $\ln(\text{O:E})_i$ the natural log of the estimated O:E ratio in the i^{th} study, and $\text{Var}\left(\ln\left(\text{O:E}\right)_i\right)$ its error variance. It is, however, also possible to use a discrete likelihood for modeling the total number of observed events in each study (O_i):

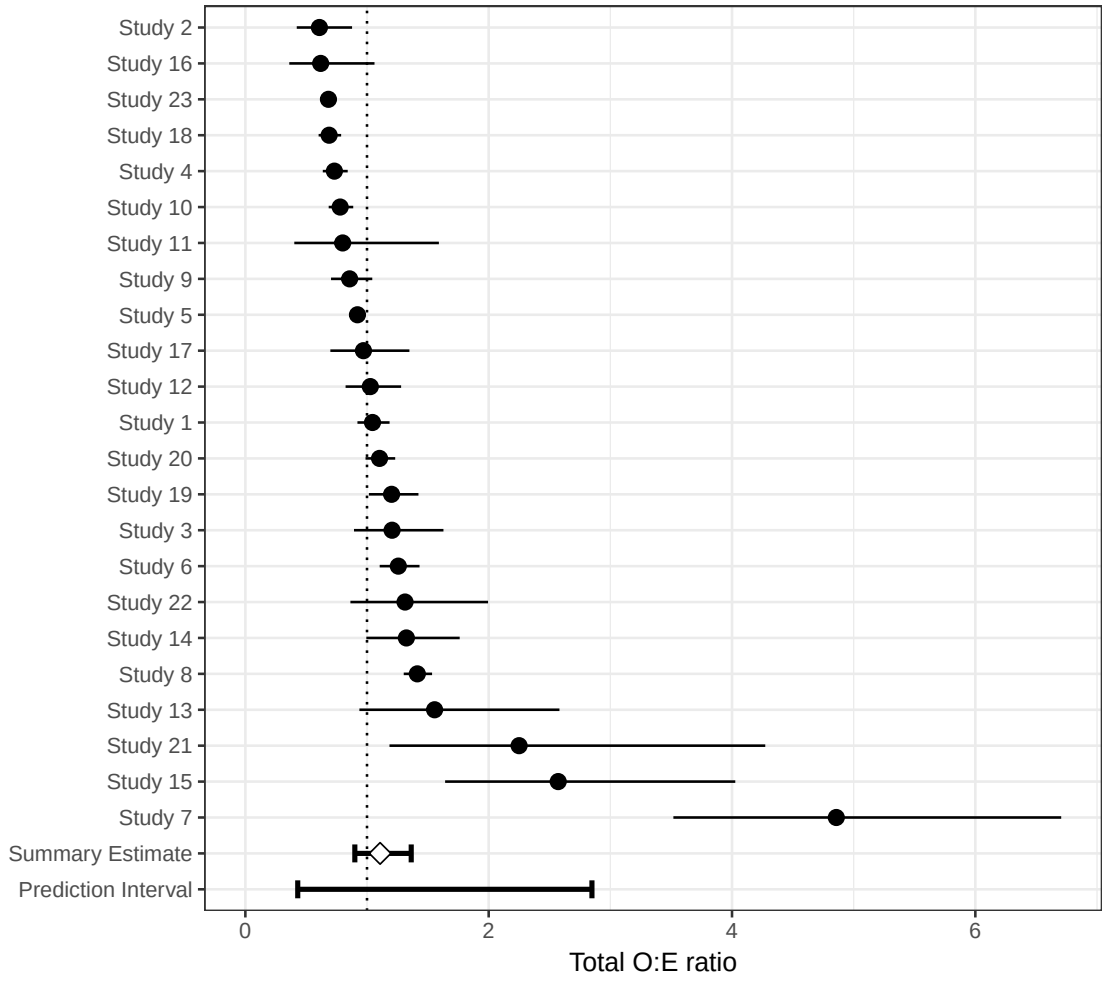
$$\begin{aligned} O_i &\sim \text{Binom}\left(N_i, p_{\text{O},i}\right) \\ \ln\left(\frac{p_{\text{O},i}}{p_{\text{E},i}}\right) &\sim \mathcal{N}\left(\mu_{\text{cal.OE}}, \tau^2_{\text{cal.OE}}\right) \end{aligned} \tag{Model 2*}$$

Alternatively, if the total sample size is unknown, we can model the within-study variation using a Poisson distribution:

$$\begin{aligned} O_i &\sim \text{Poisson}\left(E_i \exp(\eta_i)\right) \\ \eta_i &\sim \mathcal{N}\left(\mu_{\text{cal.OE}}, \tau^2_{\text{cal.OE}}\right) \end{aligned} \tag{Model 2**}$$

Note that the prior $\mu_{\text{cal.OE}} \sim \mathcal{N}(0, 10^2)$ is somewhat restricted due to exponentiation of η_i. For all models, the summary O:E ratio is simply given by $\exp(\hat{\mu}_{\text{cal.OE}})$.

Figure S11: Forest plot of the calibration performance of EuroSCORE II



For instance, the total O:E ratio of EuroSCORE II can be summarized as follows:

```
fit <- with(EuroSCORE, valmeta(measure="OE", O=n.events, E=e.events, N=n))
```

The corresponding forest plot is depicted in Figure S11. By default, Model 2 will be used for meta-analysis of the total O:E ratio. Missing information will be estimated from the available data. We can easily verify this in R :

```
> head(fit$data)
        theta   theta.se   theta.CIl   theta.CIu theta.source
1  0.04405999 0.06426711 -0.08190124  0.17002122   O, E and N
2 -0.49643689 0.18638824 -0.86175112 -0.13112265   O, E and N
3  0.18721154 0.15320840 -0.11307140  0.48749448   O, E and N
4 -0.31224698 0.07110740 -0.45161492 -0.17287904   O, E and N
5 -0.08180235 0.03603276 -0.15242525 -0.01117944   O, E and N
6  0.22897447 0.06624097  0.09914455  0.35880439   O, E and N
```

where `theta` and `theta.se` represent estimates for $\ln(\text{O:E})_i$ and, respectively, its standard error.

Since all studies provide information on the total number of observed and expected events, it is usually more appropriate to adopt a discrete within-study likelihood. We can, for instance, implement Model 2* as follows:

```
fit <- with(EuroSCORE, valmeta(measure="OE", O=n.events, E=e.events, N=n,
            pars=list(model.oe="poisson/log")))
```

Finally, when performing a Bayesian meta-analysis, we can tailor the within-study likelihood of the log O:E ratios ζ_i^*:

$$\zeta_i^* \sim \mathcal{N}\left(\mu_{\text{cal.OE}}, \tau^2_{\text{cal.OE}}\right)$$

where

- $O_i \sim \text{Binom}\left(N_i, p_{\text{O},i}\right)$ and $\zeta_i^* = \ln\left({}^{p_{\text{O},i}}/_{p_{\text{E},i}}\right)$ if O_i, $p_{\text{E},i}$ (or E_i) and N_i are reported.
- $O_i \sim \text{Poisson}\left(E_i \exp(\zeta_i^*)\right)$ if only O_i and E_i are reported.
- $\ln(\text{O:E})_i \sim \mathcal{N}\left(\zeta_i^*, \text{Var}\left(\ln(\text{O:E})_i\right)\right)$ if the total O:E ratio and its confidence interval or standard error are reported.

This approach is also known as hierarchical related regression, as different regression models are specified and linked by shared parameters (here $\mu_{\text{cal.OE}}$ and $\tau_{\text{cal.OE}}$). By default, `metamisc` assumes that $\mu_{\text{cal.OE}} \sim \mathcal{N}(0, 100)$ and $\tau_{\text{cal.OE}} \sim \text{Unif}(0, 2)$. The model can be implemented as follows:

```
fit <- with(EuroSCORE, valmeta(measure="OE", O=n.events, E=e.events, N=n,
            method="BAYES"))
```

This should yield the following results:

```
> fit
Summary O:E ratio with 95% credibility and 95% prediction interval:

 estimate       CIl       CIu       PIl       PIu
1.0975083 0.8844918 1.3381688 0.2871161 2.4691098

Penalized expected deviance:  368.14

Number of studies included:  23
```

Since all studies provide information on O_i, E_i and N_i, a binomial likelihood has been used to model all within-study variation. We can inspect this as follows:

```
> fit$runjags$model

JAGS model syntax:
```

```
1   |  model{
2   |  for (j in 1:23)
3   |  {
4   |  O[s1[j]] ~ dbinom(pobs[j], N[s1[j]])
5   |  OE[j] <- exp(theta[s1[j]])
6   |  pobs[j] <- min(OE[j], N[s1[j]]/(E[s1[j]]+1)) * E[s1[j]]/N[s1[j]]
7   |  }
8   |  for (j in 1:23)
9   |  {
10  |  theta[j] ~ dnorm(mu.logoe, bsprec.logoe)
11  |  }
12  |  bsprec.logoe <- 1/(bsTau*bsTau)
13  |  bsTau ~ dunif(0,2)
14  |  mu.logoe ~ dnorm(0,0.01)
15  |  mu.oe <- exp(mu.logoe)
16  |  pred.oe <- exp(pred.logoe)
17  |  pred.logoe ~ dnorm(mu.logoe, bsprec.logoe)
18  |  }
```

To further explore the potential advantages of Bayesian meta-analysis, we can introduce some missing values in the reported study size.

```
EuroSCORE.new <- EuroSCORE
EuroSCORE.new$n[c(1, 2, 5, 10, 20)] <-  NA
fit <- with(EuroSCORE.new, valmeta(measure="OE", O=n.events, E=e.events, N=n,
            method="BAYES"))
```

Notice that results are very similar:

```
> fit
Summary O:E ratio with 95% credibility and 95% prediction interval:

 estimate       CIl       CIu       PIl       PIu
1.0976945 0.8750624 1.3237027 0.2757592 2.4640205

Penalized expected deviance:  377.3

Number of studies included:  23
```

However, we now have a Poisson likelihood for the 5 studies where the total sample size is missing:

```
> fit$runjags$model

JAGS model syntax:

1   |  model{
2   |  for (j in 1:18)
3   |  {
4   |  O[s1[j]] ~ dbinom(pobs[j], N[s1[j]])
5   |  OE[j] <- exp(theta[s1[j]])
6   |  pobs[j] <- min(OE[j], N[s1[j]]/(E[s1[j]]+1)) * E[s1[j]]/N[s1[j]]
7   |  }
8   |  for (j in 1:5)
9   |  {
10  |  O[s2[j]] ~ dpois(lambda[j])
11  |  lambda[j] <- exp(theta[s2[j]])*E[s2[j]]
12  |  }
13  |  for (j in 1:23)
14  |  {
15  |  theta[j] ~ dnorm(mu.logoe, bsprec.logoe)
16  |  }
17  |  bsprec.logoe <- 1/(bsTau*bsTau)
18  |  bsTau ~ dunif(0,2)
```

```
19  |  mu.logoe ~ dnorm(0,0.01)
20  |  mu.oe <- exp(mu.logoe)
21  |  pred.oe <- exp(pred.logoe)
22  |  pred.logoe ~ dnorm(mu.logoe, bsprec.logoe)
23  |  }
```

4.4.3 Meta-analysis of the calibration slope

When validating a previously developed prediction model in new patient data, the calibration slope is calculated as follows:

$$\begin{aligned} y_k &\sim \text{Bernoulli}(p_k) \\ \text{logit}(p_k) &= \alpha + \beta \text{LP}_k \end{aligned}$$

where y_k is the observed (binary) outcome for individual k and LP_k is the linear predictor for that individual. The LP summarizes the effects of the predictors $\boldsymbol{X}$, and is related to the predicted probability $P(y_k = 1) = P_{\text{E},k}$, as $\text{logit}(P_{\text{E},k}) = \text{LP}_k$. The calibration slope is then simply given by β.

As discussed in the main article, the standard meta-analysis model for obtaining a summary estimate of the calibration slope is given as:

$$\begin{aligned} O_{ij} &\sim \text{Binom}(N_{ij}, p_{\text{O},ij}) \\ \text{logit}(p_{\text{O},ij}) &= \alpha_i + \beta_i \text{logit}(P_{\text{E},ij}) \\ \beta_i &\sim \mathcal{N}(\mu_{\text{cal.slope}}, \tau^2_{\text{cal.slope}}) \\ \mu_{\text{cal.slope}} &\sim \mathcal{N}(0, 10^6) \end{aligned} \qquad \text{(Model 3)}$$

with $\tau_{\text{cal.slope}} \sim \text{Unif}(0, 2)$ or $\tau_{\text{cal.slope}} \sim \text{Student-}t(0, 0.5^2, 3)\, T[0, 10]$. We use a Binomial distribution, rather than a Bernoulli distribution, to account for the fact that information on observed and expected events is only available for groups of individuals.

Note that when including validation studies with different follow-up lengths, estimates for O_{ij} and $P_{\text{E},ij}$ can be extrapolated. When operating under a Bayesian estimation framework, we can integrate uncertainty due to extrapolation in the meta-analysis model.

Suppose that t represents the time period for which calibration performance is of primary interest (e.g. $t = 10$ years for Framingham Wilson). Further, suppose that l_i represents the actual time period for which O_{ij} and E_{ij} were available in study i. For instance, when reviewing the validation studies of Framingham Wilson, some studies assessed calibration performance after $l_i = 5$ years. We then have:

$$p_t = 1 - S_{\text{KM},t} = 1 - \exp\left(\frac{t \ln(1 - p_l)}{l}\right)$$

Further, we can account for sampling error in $P_{\text{E},ij}$ by specifying a binomial distribution for E_{ij}. Model 3 then becomes:

$$\begin{aligned} O_{ij} &\sim \text{Binom}(N_{ij}, \zeta_{\text{O},ij}) \\ E_{ij} &\sim \text{Binom}(N_{ij}, \zeta_{\text{E},ij}) \\ \zeta_{\text{O},ij} &= 1 - \exp\left(\frac{l_i \ln(1 - p_{\text{O},ij})}{t}\right) \\ \zeta_{\text{E},ij} &= 1 - \exp\left(\frac{l_i \ln(1 - p_{\text{E},ij})}{t}\right) \\ \text{logit}(p_{\text{O},ij}) &= \alpha_i + \beta_i \text{logit}(p_{\text{E},ij}) \\ \text{logit}(p_{\text{E},ij}) &\sim \mathcal{N}(0, 10^6) \\ \beta_i &\sim \mathcal{N}(\mu_{\text{cal.slope}}, \tau^2_{\text{cal.slope}}) \\ \mu_{\text{cal.slope}} &\sim \mathcal{N}(0, 10^6) \end{aligned} \qquad \text{(Model 3*)}$$

with $\tau_{\text{cal.slope}} \sim \text{Unif}(0, 2)$ or $\tau_{\text{cal.slope}} \sim \text{Student-}t(0, 0.5^2, 3)\, T[0, 10]$.

As mentioned, t represents the time period for which calibration performance is of primary interest, and l_i the actual time period for which O_{ij} and E_{ij} were available in study i. For most validation studies, it is likely that $l_i = t$, in which case the Poisson extrapolation will be omitted (as $\zeta_{\mathrm{O},ij}$ then collapses to $p_{\mathrm{O},ij}$, and respectively, $\zeta_{\mathrm{E},ij}$ to $p_{\mathrm{E},ij}$).

4.4.4 Evaluation of convergence

The convergence of all estimated Bayesian meta-analysis models is verified by calculating the potential scale reduction factor (psrf) of the Gelman-Rubin statistic autocorrelation of the sample. If the psrf is greater than 1.05, a warning will be printed. For instance, consider the following model where only 100 MCMC samples are generated from the posterior distribution:

```
data(EuroSCORE)
with(EuroSCORE, valmeta(cstat=c.index, cstat.se=se.c.index,
                        cstat.95CI=cbind(c.index.95CIl, c.index.95CIu),
                        N=n, O=n.events, method="BAYES",
                        pars=list(hp.tau.dist="dhalft"), sample=100))
```

This will yield:

```
Note: Unable to calculate the multivariate psrf
Finished running the simulation
Model results for the c-statistic:

 estimate     95CIl     95CIu     95PIl     95PIu
0.7865492 0.7611015 0.8071919 0.6885138 0.8757514

Penalized expected deviance:  83.7

Number of studies included:  23
Note: For 4 validation(s), the standard error of the concordance statistic was estimated
using method 'Newcombe.4'.
Warning message:
In print.valmeta(x) :
  Model did not properly converge! The upper bound of the convergence diagnostic (psrf)
  exceeds 1.05 for the parameters bsTau (psrf=1.68) . Consider re-running the analysis
  by increasing the optional arguments 'adapt', 'burnin' and/or 'sample'.
```

4.5 Performance in sparse data

We conducted a small simulation study to investigate the performance of frequentist and Bayesian estimation methods in sparse data. Hereto, we assessed to what extent estimates for the between-study standard deviation are affected when few studies are available for meta-analysis.

We first performed a meta-analysis of 23 studies reporting the discrimination and calibration performance of EuroSCORE II. We also performed a meta-analysis of the discrimination ($K = 21$) and calibration ($K = 17$) performance of the Framingham Risk Score. Results in Table S8 indicate that estimates for the between-study standard deviation (defined as τ_{discr} and $\tau_{\text{cal.OE}}$) are fairly stable, regardless the estimation method. Hence, we will use these as reference values in our simulation study.

Table S8: Meta-analysis estimates from the empirical examples

Example	**Estimation**	τ_{discr}			$\tau_{\text{cal.OE}}$		
		Estimate	2.5% CI	97.5% CI	Estimate	2.5% CI	97.5% CI
ES2	ML	0.2557	0.1753	0.4226	0.4323	0.3325	0.6696
	REML	0.2647	0.1753	0.4226	0.4444	0.3325	0.6696
	Bayesian †	0.2792	0.1736	0.4178	0.4541	0.3057	0.6386
	Bayesian ‡	0.2725	0.1684	0.4034	0.4521	0.2958	0.6318
FRS	ML	0.1917	0.1283	0.2917	0.5110	0.3862	0.8072
	REML	0.1975	0.1283	0.2917	0.5272	0.3862	0.8072
	Bayesian †	0.2074	0.1348	0.3083	0.5553	0.3711	0.8081
	Bayesian ‡	0.2053	0.1306	0.2986	0.5519	0.3761	0.7975

ES2 = EuroSCORE II; FRS = Framingham Risk Score; ML = Maximum Likelihood; REML = Restricted Maximum Likelihood; CI = confidence (for ML and REML) or credibility (for Bayesian meta-analysis) interval. For meta-analysis of calibration performance, we assumed within-study Normality for the log O:E ratio.

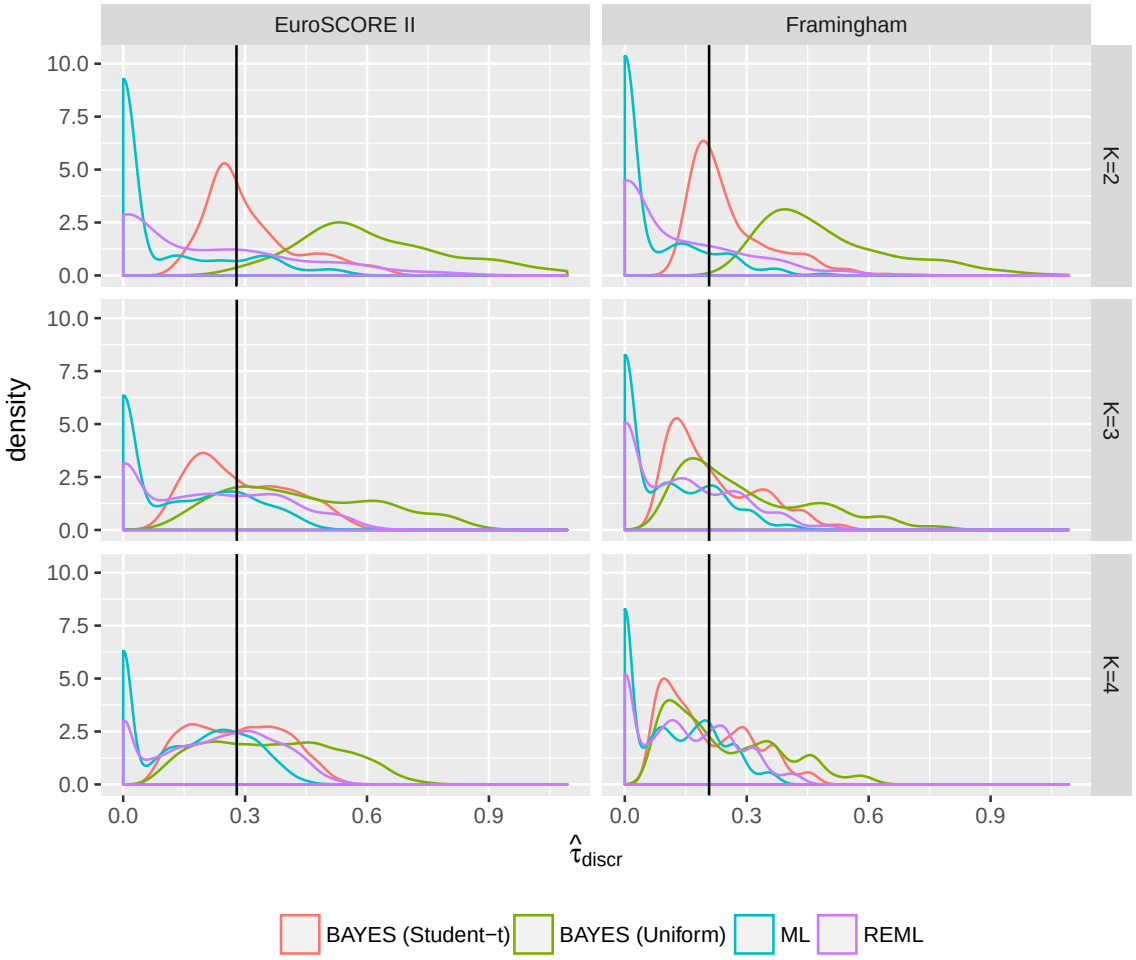
† It was assumed that $\tau_{\text{discr}} \sim \text{Unif}(0, 2)$ and $\tau_{\text{discr}}, \tau_{\text{cal.OE}} \sim \text{Unif}(0, 2)$

‡ It was assumed that $\tau_{\text{discr}} \sim \text{Student-}t(0, 0.5^2, 3)\,T[0, 10]$ and $\tau_{\text{cal.OE}} \sim \text{Student-}t(0, 1.5^2, 3)\,T[0, 10]$

We adopted a full factorial design to form unique study subsets for meta-analysis. For EuroSCORE II, this approach allows to perform a total of $23!/(2!(23-2)!) = 253$ meta-analyses that are based on $K = 2$ studies. Similarly, a total of 1771 and 8855 meta-analyses can be performed for $K = 3$ and, respectively, $K = 4$. The same approach was used for FRS, yielding 210 (for $K = 2$), 1330 (for $K = 3$) and 5985 (for $K = 4$) estimates for τ_{discr}.

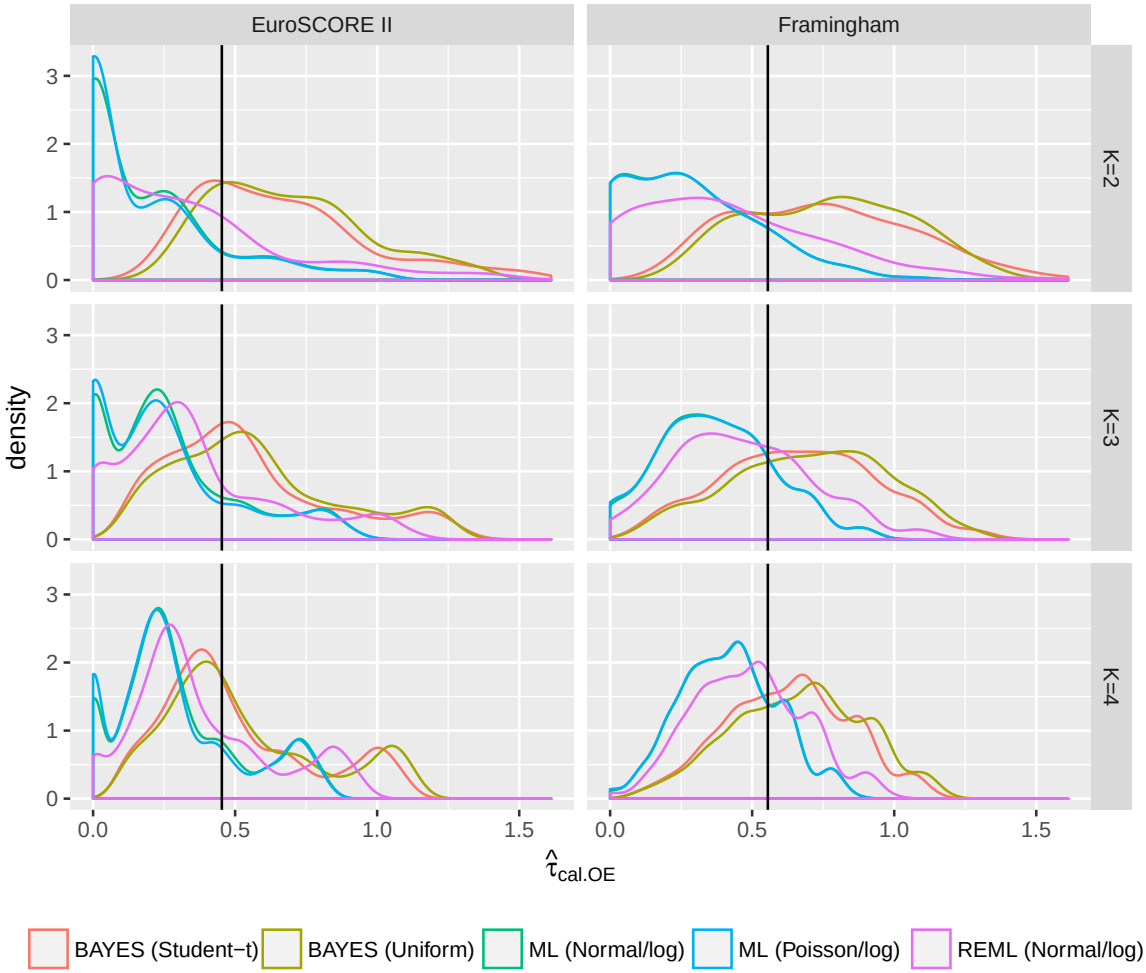
Results in Figure S12 indicate that (restricted) maximum likelihood estimation tends to underestimate the presence of between-study heterogeneity when meta-analyses are based on few studies. Conversely, when adopting Bayesian estimation with the proposed prior distributions, estimates of τ_{discr} are much closer to the reference values, particularly when using a Student-t prior.

Figure S12: Estimates of τ_{discr}



Results for meta-analyses that are based on K validation studies. For Bayesian meta-analysis models, $\hat{\tau}_{\text{discr}}$ represents the posterior median. The vertical line indicates the posterior median of the Bayesian meta-analysis model using a uniform prior in the full set of 23 (for EuroSCORE II) or 21 (for the FRS) studies.

Figure S13: Estimates of $\tau_{\text{cal.OE}}$



Results for meta-analyses that are based on K validation studies. For Bayesian meta-analysis models, $\hat{\tau}_{\text{cal.OE}}$ represents the posterior median. The vertical line indicates the posterior median of the Bayesian meta-analysis model using a uniform prior in the full set of 23 (for EuroSCORE II) or 17 (for the FRS) studies. For each graph, $23!/(K!(23-K)!)$ and, respectively, $17!/(K!(17]-K)!)$ meta-analyses were performed for EuroSCORE II and the Framingham Risk Score.

5 Additional Results

5.1 Meta-analysis of EuroSCORE II

Table S9: Meta-analysis estimates for the EuroSCORE II model

Performance	**Estimation**	**Model**	K	$\hat{\mu}$	**(95% CI)**	$\hat{\tau}$	**(95% CI)**
c-statistic	REML	Model 1	23	1.32	(1.18 ; 1.46)	0.26	(0.18 ; 0.42)
	Bayesian†	Model 1	23	1.32	(1.17 ; 1.46)	0.28	(0.18 ; 0.42)
	Bayesian‡	Model 1	23	1.32	(1.17 ; 1.45)	0.27	(0.17 ; 0.40)
Total O:E ratio	REML	Model 2	23	0.10	(-0.11 ; 0.31)	0.44	(0.33 ; 0.67)
	Bayesian†	Model 2*	23	0.09	(-0.12 ; 0.30)	0.45	(0.31 ; 0.64)
	Bayesian‡	Model 2*	23	0.09	(-0.12 ; 0.29)	0.45	(0.30 ; 0.64)
	ML$^{\bullet}$	Model 2**	23	0.09	(-0.10 ; 0.29)	0.42	(0.24 ; 0.56)
	Bayesian†	Model 2**	23	0.09	(-0.12 ; 0.29)	0.45	(0.30 ; 0.64)
	Bayesian‡	Model 2**	23	0.09	(-0.12 ; 0.29)	0.45	(0.30 ; 0.63)

K = 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

† 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

$^{\bullet}$ Confidence intervals were approximated using parametric bootstrapping.

5.2 Meta-analysis of the Framingham Risk Score

As a sensitivity analysis, we conducted a meta-analysis of calibration performance where we omitted studies with inappropriate follow-up. We also conducted a meta-analysis where observed and expected event rates were extrapolated using a Poisson distribution.

For the calibration slope, we found that inclusion of validation studies with calibration performance reported for 5 or 7.5 years follow-up facilitated estimation of the meta-analysis models. In particular, this strategy helped to increase the precision of the summary estimate, and to identify $\tau_{\text{cal.slope}}$ (which was 0 for $K = 3$, but increased to 0.2 when estimated using a Bayesian framework with $K = 11$).

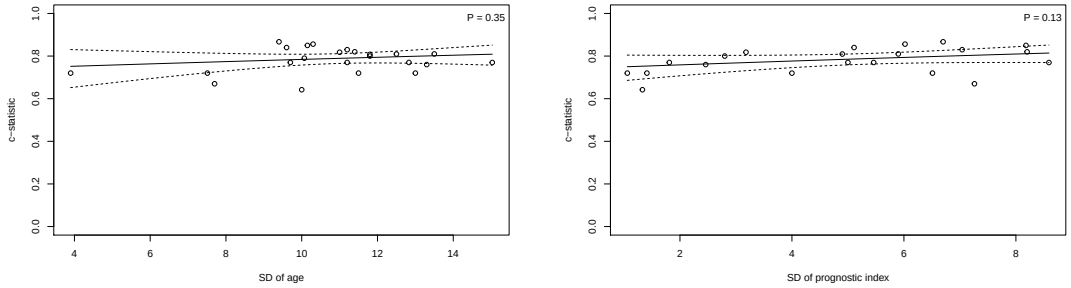
Table S10: Meta-analysis estimates for the Framingham Risk Score

Performance	**Estimation**	**Model**	K	**Summary**	**95% CI**	**95% PI**	
Total O:E ratio	REML	Model 2	6	0.56	0.28 – 1.16	0.09 – 3.62	
	Bayesian†	Model 2	6	0.61	0.19 – 1.08	0.00 – 2.84	
	Bayesian‡	Model 2	6	0.61	0.20 – 1.07	0.00 – 2.63	
	ML	Model 2*	6	0.56	0.25 – 1.26	0.03 – 11.29	$\star$
	Bayesian†	Model 2*	7	0.60	0.19 – 1.09	0.00 – 2.91	
	Bayesian‡	Model 2**	7	0.60	0.18 – 1.05	0.00 – 2.67	
	REML$^{\bullet}$	Model 2	16	0.58	0.44 – 0.76	0.19 – 1.72	
	Bayesian$^{\dagger\bullet}$	Model 2	16	0.58	0.42 – 0.75	0.09 – 1.49	
	Bayesian$^{\ddagger\bullet}$	Model 2	16	0.58	0.42 – 0.75	0.09 – 1.48	
	ML$^{\bullet}$	Model 2**	16	0.57	0.35 – 0.94	0.06 – 5.25	$\star$
	Bayesian$^{\dagger\bullet}$	Model 2**	17	0.58	0.43 – 0.75	0.09 – 1.46	
	Bayesian$^{\ddagger\bullet}$	Model 2**	17	0.58	0.42 – 0.74	0.10 – 1.48	
Calibration slope	ML	Model 3	3	1.03	0.90 – 1.16	0.20 – 1.87	
	Bayesian†	Model 3	3	1.05	0.47 – 1.64	-0.01 – 2.22	
	Bayesian‡	Model 3	3	1.05	0.51 – 1.65	-0.06 – 2.17	
	Bayesian$^{\dagger\bullet}$	Model 3*	11	0.98	0.80 – 1.17	0.48 – 1.51	
	Bayesian$^{\ddagger\bullet}$	Model 3*	11	0.99	0.81 – 1.16	0.51 – 1.47	

Summary estimates for calibration performance (Total O:E ratio and calibration slope) were derived for a time period of 10 years. Where necessary, observed and expected event rates were extrapolated using a Poisson distribution ($^{\bullet}$).
K = 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
$\star$ statistical model not converged

5.3 Meta-regression EuroSCORE II

Figure S14: Results from random-effects meta-regression models for EuroSCORE II. Full lines indicate the bounds of the 95% confidence interval around the regression line. Dots indicate the included validation studies.



References

[1] Nashef SAM, Roques F, Sharples LD et al. EuroSCORE II ; 41(4): 734–744; discussion 744–745. DOI:10.1093/ejcts/ezs043. 22378855.

[2] Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: Asymptotic methods and evaluation ; 25(4): 559–573. DOI:10.1002/sim.2324. 16217835.

[3] Wilson PW, D'Agostino RB, Levy D et al. Prediction of coronary heart disease using risk factor categories ; 97(18): 1837–1847. DOI:10.1161/âĂŃ01.CIR.97.18.1837. 9603539.

[4] Damen JAAG, Hooft L, Schuit E et al. Prediction models for cardiovascular disease risk in the general population: Systematic review ; 353: i2416. DOI:10.1136/bmj.i2416.

[5] Damen JAAG, Pajouheshnia R, Heus P et al. Performance of the Framingham risk models and Pooled Cohort Equations: A systematic review and meta-analysis ; Submitted.

[6] Buitrago F, Calvo-Hueros JI, Cañón-Barroso L et al. Original and REGICOR Framingham functions in a nondiabetic population of a Spanish health care center: A validation study 2011 Sep-Oct; 9(5): 431–438. DOI:10.1370/afm.1287. 21911762.

[7] Comín E, Solanas P, Cabezas C et al. [estimating cardiovascular risk in spain using different algorithms] ; 60(7): 693–702. 17663853.

[8] D'Agostino RB, Grundy S, Sullivan LM et al. Validation of the Framingham coronary heart disease prediction scores: Results of a multiple ethnic groups investigation ; 286(2): 180–187. 11448281.

[9] DeFilippis AP, Young R, Carrubba CJ et al. An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort ; 162(4): 266–275. DOI:10.7326/M14-1281. 25686167.

[10] Empana JP, Ducimetière P, Arveiler D et al. Are the Framingham and PROCAM coronary heart disease risk functions applicable to different European populations? The PRIME Study ; 24(21): 1903–1911. 14585248.

[11] Ferrario M, Chiodini P, Chambless LE et al. Prediction of coronary events in a low incidence population. Assessing accuracy of the CUORE Cohort Study prediction equation ; 34(2): 413–421. DOI:10.1093/ije/dyh405. 15659467.

[12] Jee SH, Jang Y, Oh DJ et al. A coronary heart disease prediction model: The Korean Heart Study ; 4(5): e005025. DOI:10.1136/bmjopen-2014-005025. 24848088.

[13] Lloyd-Jones DM, Wilson PWF, Larson MG et al. Framingham risk score and prediction of lifetime risk for coronary heart disease ; 94(1): 20–24. DOI:10.1016/j.amjcard.2004.03.023. 15219502.

[14] Mainous AG, Koopman RJ, Diaz VA et al. A coronary heart disease risk score based on patient-reported information ; 99(9): 1236–1241. DOI:10.1016/j.amjcard.2006.12.035. 17478150.

[15] Marrugat J, Subirana I, Comín E et al. Validity of an adaptation of the Framingham cardiovascular risk function: The VERIFICA Study ; 61(1): 40–47. DOI:10.1136/jech.2005.038505. 17183014.

[16] Reissigová J and Zvárová J. The Framingham risk function underestimated absolute coronary heart disease risk in Czech men ; 46(1): 43–49. 17224979.

[17] Rodondi N, Locatelli I, Aujesky D et al. Framingham risk score and alternatives for prediction of coronary heart disease in older adults ; 7(3): e34287. DOI:10.1371/journal.pone.0034287. 22470551.

[18] Ryckman EM, Summers RM, Liu J et al. Visceral fat quantification in asymptomatic adults using abdominal CT: Is it predictive of future cardiac events? ; 40(1): 222–226. DOI:10.1007/s00261-014-0192-z. 25015400.

[19] Simmons RK, Sharp S, Boekholdt SM et al. Evaluation of the Framingham risk score in the European Prospective Investigation of Cancer-Norfolk cohort: Does adding glycated hemoglobin

improve the prediction of coronary heart disease events? ; 168(11): 1209–1216. DOI:10.1001/archinte.168.11.1209. 18541829.

[20] Suka M, Sugimori H and Yoshida K. Application of the updated Framingham risk score to Japanese men ; 24(6): 685–689. 11768728.

[21] Vaidya D, Yanek LR, Moy TF et al. Incidence of coronary artery disease in siblings of patients with premature coronary artery disease: 10 years of follow-up ; 100(9): 1410–1415. DOI:10.1016/j.amjcard.2007.06.031. 17950799.

[22] Hanley JA and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve ; 143(1): 29–36. DOI:10.1148/radiology.143.1.7063747. 7063747.

[23] Carosella V, Mastantuono C, Golovonevsky V et al. Prospective and multicentric validation of the ArgenSCORE in aortic valve replacement surgery. Comparison with the EuroSCORE I and the EuroSCORE II ; 82: 6–12. Carosella VC, Mastantuono C, Golovonevsky V, Cohen V, Grancelli H, Rodriguez W, et al. . 2014;82:6-12.

[24] Borde D, Gandhe U, Hargave N et al. The application of European system for cardiac operative risk evaluation II (EuroSCORE II) and Society of Thoracic Surgeons (STS) risk-score for risk stratification in Indian patients undergoing cardiac surgery 2013 Jul-Sep; 16(3): 163–166. DOI:10.4103/0971-9784.114234. 23816669.

[25] White IR, Rapsomaniki E and Emerging Risk Factors Collaboration. Covariate-adjusted measures of discrimination for survival data ; 57(4): 592–613. DOI:10.1002/bimj.201400061. 25530064.

[26] Austin PC and Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: Relation to the variance and odds ratio of a continuous explanatory variable ; 12: 82. DOI:10.1186/1471-2288-12-82. 22716998.

[27] Debray TPA, Damen JAAG, Snell KIE et al. A guide to systematic review and meta-analysis of prediction model performance ; 356: i6460. DOI:10.1136/bmj.i6460.

[28] Tierney JF, Stewart LA, Ghersi D et al. Practical methods for incorporating summary time-to-event data into meta-analysis ; 8: 16. DOI:10.1186/1745-6215-8-16. 17555582.

[29] Parmar MKB, Torri V and Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints ; 17(24): 2815–2834. DOI:10.1002/(SICI)1097-0258(19981230)17:24<2815::AID-SIM110>3.0.CO;2-8.

[30] Kengne AP, Beulens JWJ, Peelen LM et al. Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): A validation of existing models ; 2(1): 19–29. DOI:10.1016/S2213-8587(13)70103-7. 24622666.

[31] Thompson DD, Murray GD, Dennis M et al. Formal and informal prediction of recurrent stroke and myocardial infarction after stroke: A systematic review and evaluation of clinical prediction models in a new cohort ; 12: 58. DOI:10.1186/1741-7015-12-58. 24708686.

[32] Ford MK, Beattie WS and Wijeysundera DN. Systematic review: Prediction of perioperative cardiac complications and mortality by the revised cardiac risk index ; 152(1): 26–35. DOI:10.7326/0003-4819-152-1-201001050-00007. 20048269.

[33] Snell K, Hua H, Debray T et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model ; 69: 40–50. DOI:10.1016/j.jclinepi.2015.05.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0895435615002309.

[34] Gagné M, Moore L, Beaudoin C et al. Performance of International Classification of Diseases-based injury severity measures used to predict in-hospital mortality: A systematic review and meta-analysis ; 80(3): 419–426. DOI:10.1097/TA.0000000000000944. 26713976.

[35] Guida P, Mastro F, Scrascia G et al. Performance of the European System for Cardiac Operative Risk Evaluation II: A meta-analysis of 22 studies involving 145,592 cardiac surgery procedures ; 148(6): 3049–3057.e1. DOI:10.1016/j.jtcvs.2014.07.039. 25161130.

[36] Meads C, Ahmed I and Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance ; 132(2): 365–377. DOI:10.1007/s10549-011-1818-2. 22037780.

[37] Chalmers JD, Mandal P, Singanayagam A et al. Severity assessment tools to guide ICU admission in community-acquired pneumonia: Systematic review and meta-analysis ; 37(9): 1409–1420. DOI:10.1007/s00134-011-2261-x. 21660535.

[38] Shen JH, Chen HL, Chen JR et al. Comparison of the Wells score with the revised Geneva score for assessing suspected pulmonary embolism: A systematic review and meta-analysis ; 41(3): 482–492. DOI:10.1007/s11239-015-1250-2. 26178041.

[39] Marques A, Ferreira RJO, Santos E et al. The accuracy of osteoporotic fracture risk prediction tools: A systematic review and meta-analysis ; 74(11): 1958–1967. DOI:10.1136/annrheumdis-2015-207907. 26248637.

[40] van Klaveren D, Steyerberg EW, Perel P et al. Assessing discriminative ability of risk models in clustered data ; 14: 5. DOI:10.1186/1471-2288-14-5. 24423445.

[41] Ohle R, O'Reilly F, O'Brien KK et al. The Alvarado score for predicting acute appendicitis: A systematic review ; 9: 139. DOI:10.1186/1741-7015-9-139. 22204638.

[42] Zhu W, He W, Guo L et al. The HAS-BLED Score for Predicting Major Bleeding Risk in Anticoagulated Patients With Atrial Fibrillation: A Systematic Review and Meta-analysis ; 38(9): 555–561. DOI:10.1002/clc.22435. 26418409.

[43] Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper) ; 1(3): 515–534. DOI:10.1214/06-BA117A.

[44] Spiegelhalter D. *Prior Distributions*. John Wiley & Sons, Ltd. ISBN 978-0-470-09260-6 978-0-471-49975-6. pp. 139–180. DOI:10.1002/0470092602.ch54.