UMC Utrecht
Julius Center

Systematic reviews of prognostic studies IV

meta-analytical approaches in systematic reviews of prognostic studies

Thomas Debray, Johanna Damen, Karel Moons
for the Cochrane Prognosis Review Methods Group

Keele University
Cochrane
University Medical Center Utrecht

Conflict of interest

We have no actual or potential conflict of interest in relation to this presentation

Basic & Advanced Courses

MSC EPIDEMIOLOGY

Systematic Reviews, Diagnostic Research, Prognostic Research, Clinical Trials, Clinical Epidemiology, Statistics

Face2Face + Online accessible from all over the world

For example:	Start date
Introduction to Statistics	18 September 2017
Systematic Reviews of Diagnostic Studies	23 October 2017
Clinical Epidemiology	30 October 2017
Systematic Reviews of Prognostic Studies	20 November 2017
Systematic Reviews of Intervention Studies	12 March 2018
Advanced Diagnostic Research	2 April 2018
Prognostic Research	25 June 2018

Face-to-face courses: www.msc-epidemiology.nl

Online courses: www.msc-epidemiology.online

UMC Utrecht Julius Center

Universiteit Utrecht

elevate

Prediction

- Risk prediction = foreseeing / foretelling
  … (probability) of something that is yet unknown

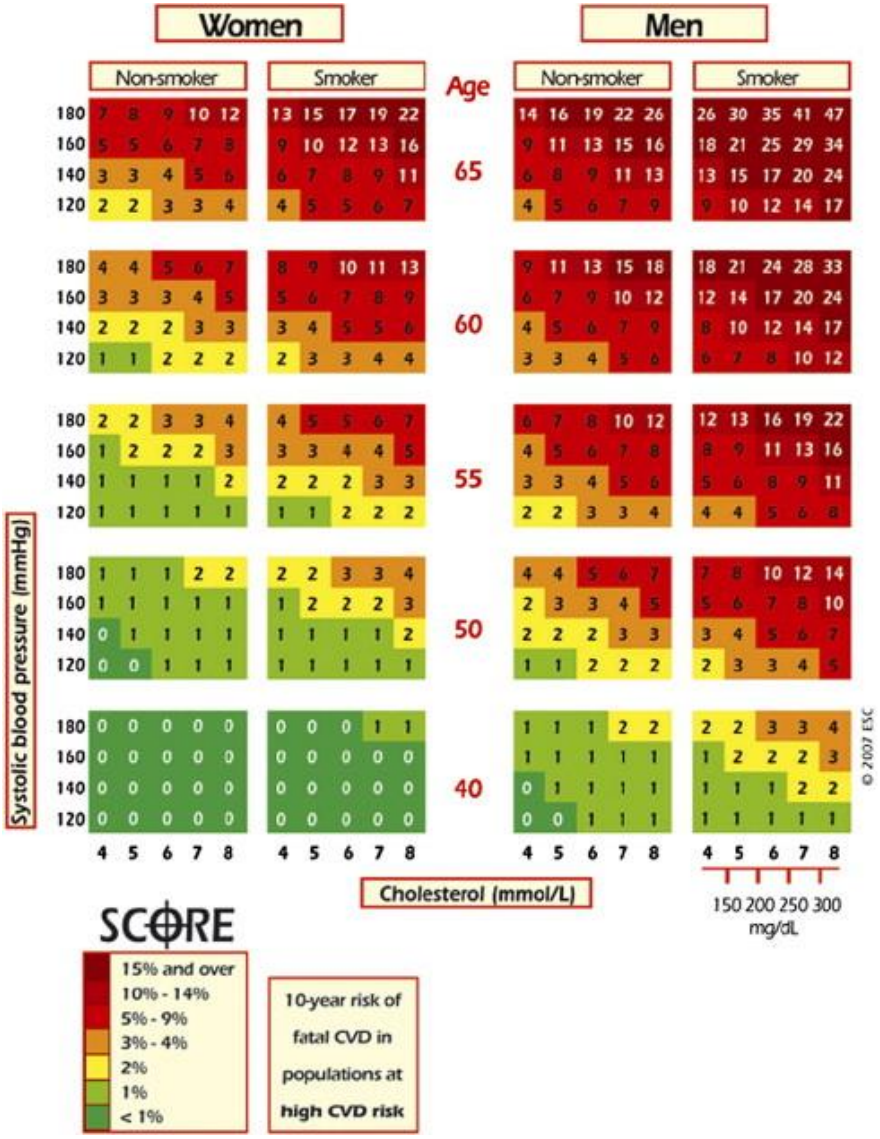
- Turn available information (predictors) into a statement about the probability:
  … diagnosis
  … prognosis

What is the big difference between diagnostic and prognostic ‘prediction’?

Prediction models

APGAR
Test Scoring

	Score 0	Score 1	Score 2
Appearance	Blue all over	Blue only at extremities	No blue coloration
Pulse	No pulse	<100 beats/min.	>100 beats/min.
Grimace	No response to stimulation	Grimace or feeble cry when stimulated	Sneezing, coughing, or pulling away when stimulated
Activity	No movement	Some movement	Active movement
Respiration	No breathing	Weak, slow, or irregular breathing	Strong cry



Three phases of Prediction Modelling

BMJ series 2009 (Altman, Moons, Royston, Vergouwe)

1. Developing a prediction model
2. Validate (+update) the model in other subjects
3. Quantify model's impact on doctor's decision making and patient outcome (cost-effectiveness)

What is big difference between 3 versus 1-2?

Focus on 1-2

Validation of prediction models

UMC Utrecht

Recap: what are validation studies?

- Apply the CPM to new individuals
  - Internal validation
  - Temporal validation
  - Geographical validation
  - Domain validation

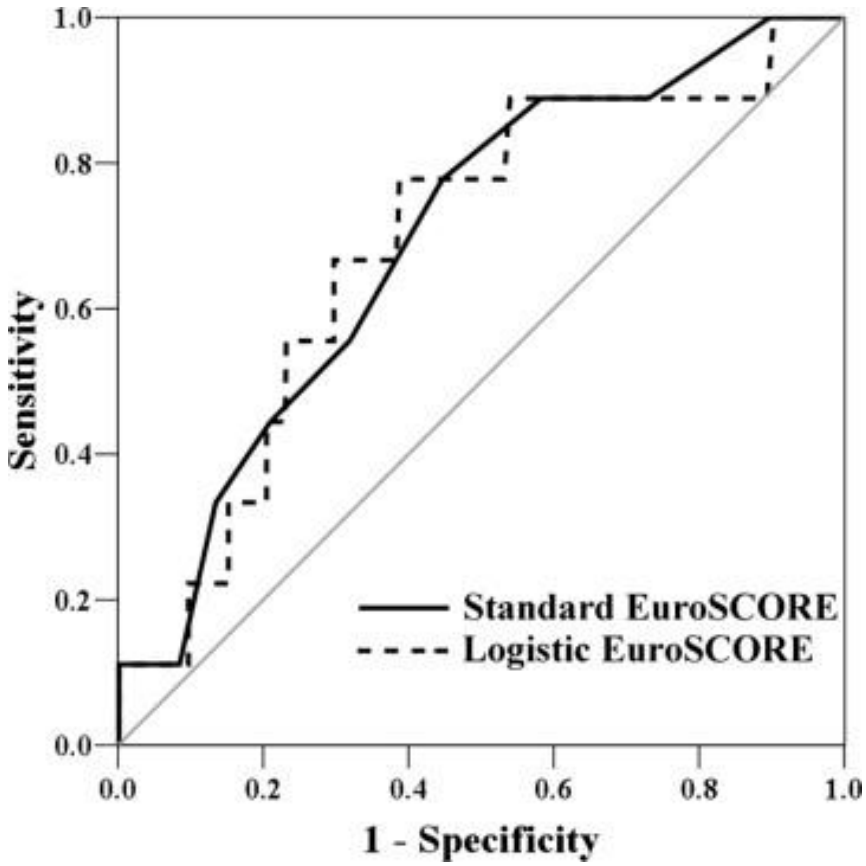
- Evaluate the predictive accuracy
  - Overall performance
  - Calibration
  - Discrimination

Performance measures

- Overall performance
  - R^2
- Discrimination
  - C-statistic, area under the ROC curve
  - Discrimination Index
- Calibration
  - Ratio of observed and expected events
  - Calibration-in-the-large
  - Calibration slope

Discrimination

ROC curve



What c-statistic does the ROC curve indicate?

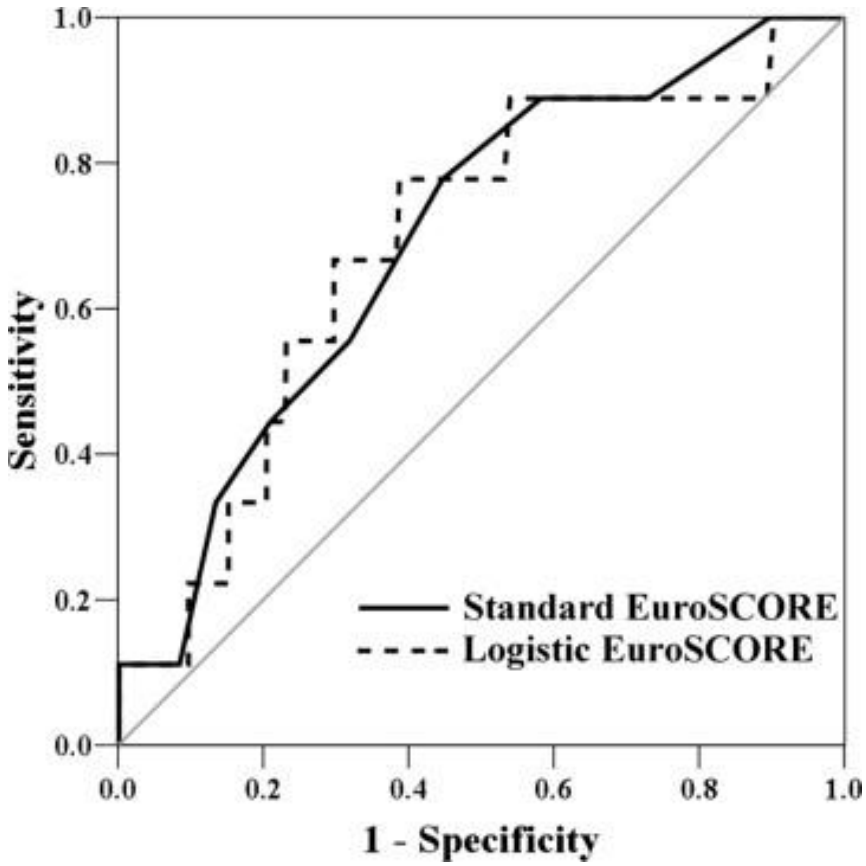
(a) 0.75 – 1.00

(b) 0.60 – 0.75

(c) < 0.60

Discrimination

ROC curve



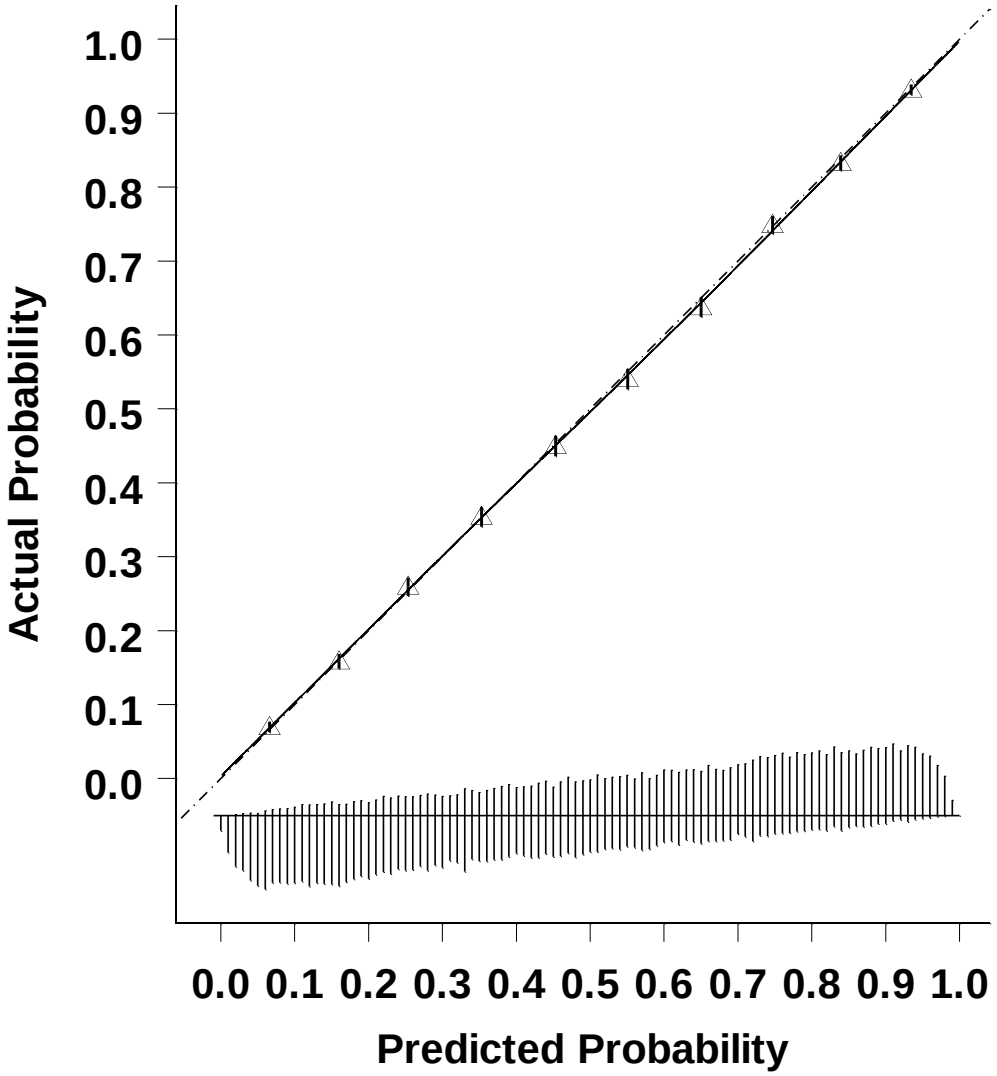
What c-statistic does the ROC curve indicate?

(a) 0.75 – 1.00

(b) 0.60 – 0.75 (0.71)

(c) < 0.60

Calibration plot – good model?

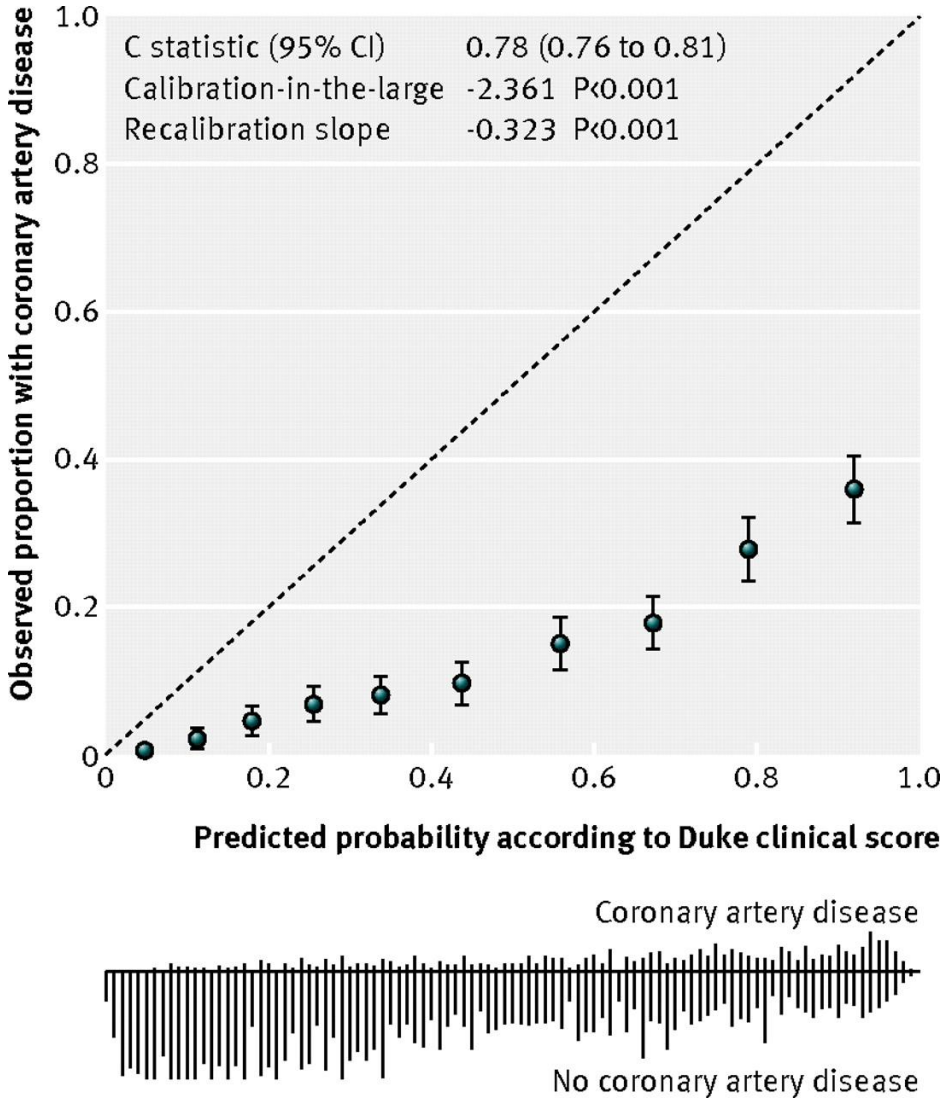


Ideal calibration

O:E = 1

Slope = 1

Calibration plot – good model?



Ref: Genders et al. Prediction model to estimate presence of coronary artery disease: retrospective pooled analysis of existing cohorts. BMJ 2012

Calibration table – good model?

External validation of EuroSCORE

Expected mortality (%) versus observed in-hospital mortality

Score	N	Expected	Observed
0-2	201	1.4	0.5
3-5	309	4.0	1.0
6-8	181	6.8	2.2
>= 9	66	10.5	3.0

How well does the EuroSCORE calibrate?

(a) Good

(b) Poor, due to over-prediction

(c) Poor, due to under-prediction

Calibration table – good model?

External validation of EuroSCORE

Expected mortality (%) versus observed in-hospital mortality

Score	N	Expected	Observed
0-2	201	1.4	0.5
3-5	309	4.0	1.0
6-8	181	6.8	2.2
>= 9	66	10.5	3.0

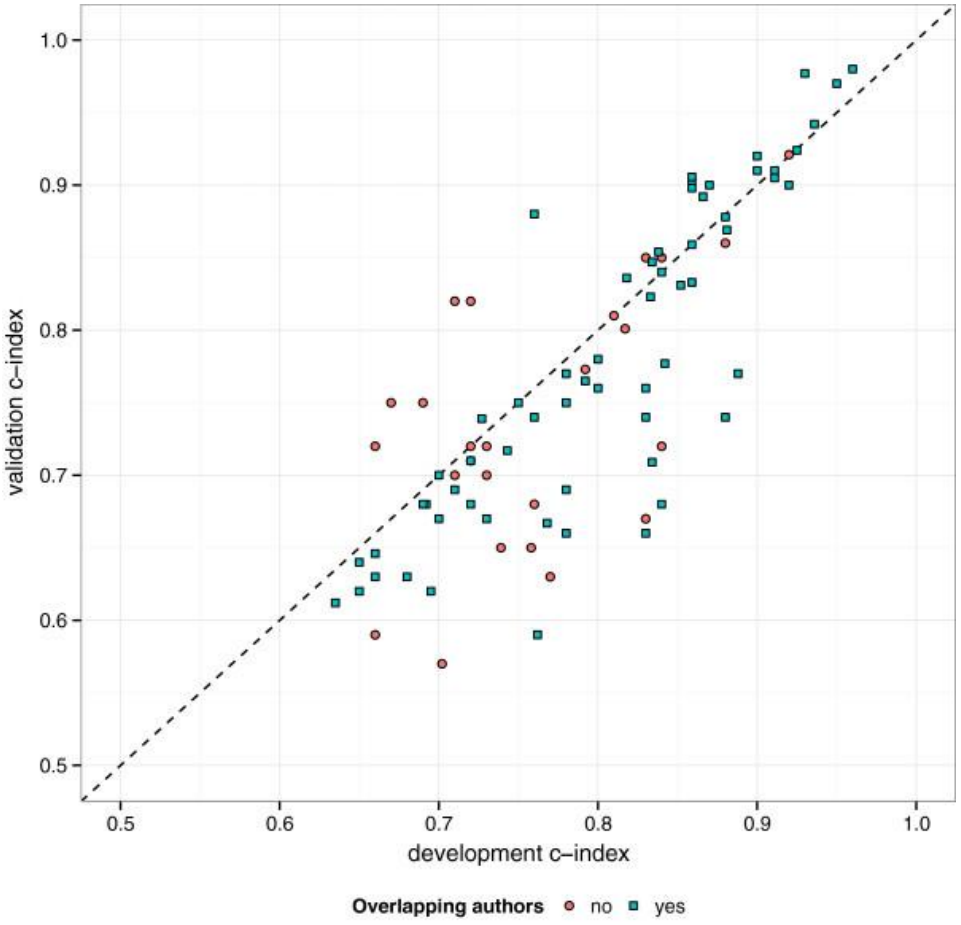
How well does the EuroSCORE calibrate?

(a) Good

(b) Poor, due to over-prediction

(c) Poor, due to under-prediction

Overfitting



Model performance often over-optimistic in the development sample

Ref: Collins et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Meth 2014

Synthesis of validation studies

UMC Utrecht

How informative are validation studies?

A single validation study may provide some information about

- **Reproducibility**: does the model perform well in new subjects from the same population? (~overfitting)
- **Transportability**: does the model perform well in new settings or populations?

Multiple model validations are usually needed to identify model generalizability across different settings & populations

We need systematic review and meta-analysis of validation studies

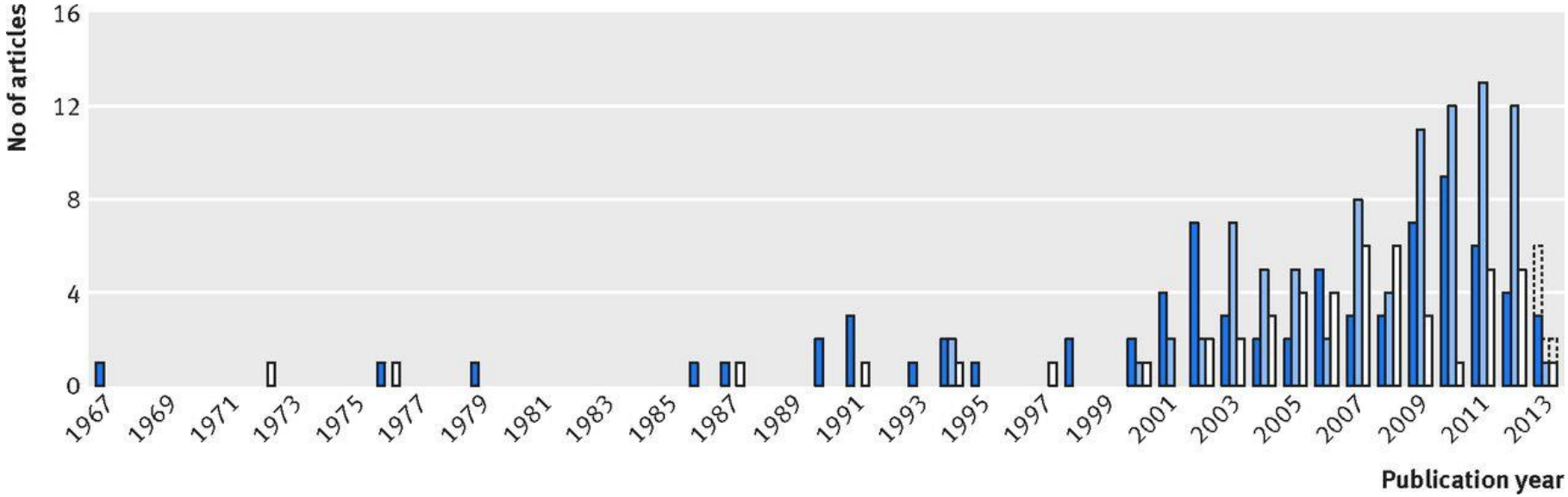
Aims

- Summarize model performance
- Investigate generalizability

Prediction models for cardiovascular disease

- Cardiovascular disease major disease burden
- Decide which people need treatment to lower risk
  - Antihypertensive medication
  - Lipid lowering medication
  - Lifestyle interventions
- Prediction models used for risk stratification
- Excess of prediction models in various fields, **with numerous validations**

Prediction models for cardiovascular disease



Development of CPM (dark blue), Validation of CPM (light blue), development + validation of CPM (white)

Johanna A A G Damen et al. BMJ 2016;353:bmj.i2416

thebmj

Prediction models for cardiovascular disease

Top 10 validated models	N
Framingham Wilson 1998	80
Framingham Anderson 1991 Am H J	73
SCORE Conroy 2003	63
Framingham D'Agostino 2008	44
Framingham unreferenced	32
Framingham ATP III 2002	31
Framingham Anderson 1991 Circulation	30
QRISK Hippisley-Cox 2007	12
PROCAM Assman 2002	8
Framingham Wolf 1991	8

Prediction models for cardiovascular disease

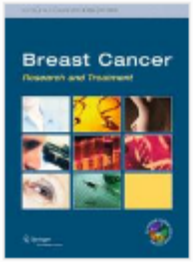
Characteristics	Framingham		SCORE: Conroy 2003[6] (n=63)
	Wilson 1998[5] (n=89)†	Anderson 1991[3] (n=73)	
Location:			
Asia	9 (10)	3 (4)	2 (3)
Australia	0 (0)	12 (16)	4 (6)
Europe	34 (38)	52 (71)	47 (75)
North America	46 (52)	6 (8)	10 (16)
Age:			
Same age range as development study*	2 (3)	21 (29)	4 (6)
Young people (<50 years)	3 (3)	6 (8)	4 (6)
Older people (>60 years)	5 (6)	7 (10)	4 (6)
Other	79 (89)	39 (53)	51 (81)
Sex:			
Men	38 (43)	30 (41)	23 (37)
Women	29 (33)	25 (34)	23 (37)
Men and women	22 (25)	18 (25)	17 (27)
Median (range) No of participants	2716 (100-163 627), n=87	2423 (262-797 373), n=71	8025 (262-44 649), n=63
Median (range) No of events	146 (8-24 659), n=65	128 (1-42 408), n=59	224 (16-1722), n=54
Median (range) C statistic	0.71 (0.57-0.92), n=61	0.75 (0.53-0.99), n=46	0.75 (0.62-0.91), n=28
Median (range) observed:expected	0.59 (0.37-1.92), n=14	0.68 (0.18-2.60), n=42	0.68 (0.28-1.50), n=26

Prediction models for cardiovascular disease

Poor and inconsistent reporting of prediction model performance.

- Poor study design
- Inappropriate handling and acknowledgement of missing data
- Calibration often omitted from the publication

Meta-analysis: is it even possible?



Breast Cancer Research and Treatment
April 2012, Volume 132, Issue 2, pp 365–377

A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance

Authors | Authors and affiliations

Catherine Meads, Ikhlaaq Ahmed, Richard D. Riley

Review

First Online: 22 October 2011

DOI: 10.1007/s10549-011-1818-2

Cite this article as:
Meads, C., Ahmed, I. & Riley, R.D.
Breast Cancer Res Treat (2012) 132:
365. doi:10.1007/s10549-011-1818-2

Meta-analysis: is it even possible?

Guidance

thebmj

RESEARCH METHODS AND REPORTING

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

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

Ref: BMJ 2017; 356 doi: https://doi.org/10.1136/bmj.i6460 (Published 05 January 2017)

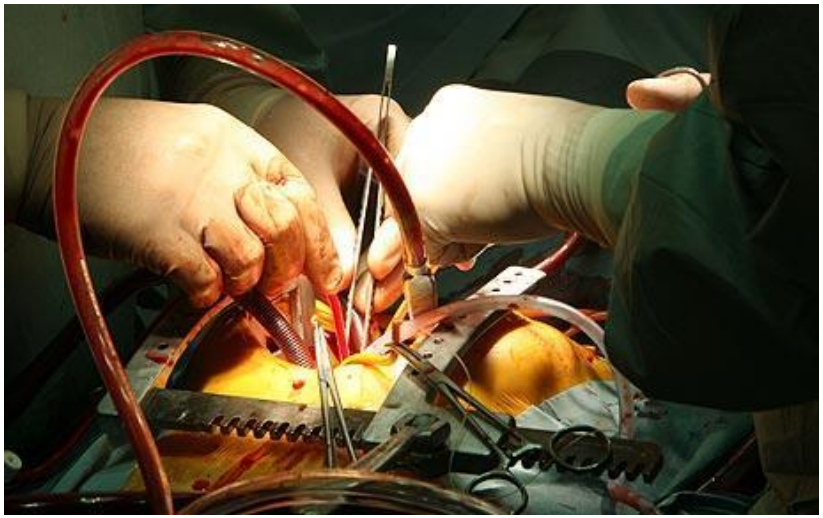
Recommended steps

1. Formulating the review question
2. Formulating the search strategy
3. Critical appraisal (CHARMS & PROBAST)
4. Quantitive data extraction
5. Meta-analysis
6. Investigating heterogeneity across studies
7. Sensitivity analyses
8. Reporting (TRIPOD)

Illustrative example
EuroSCORE

Predicting mortality after cardiac surgery

- Cardiac surgery in high-risk population
- Need for risk stratification
- Establish risk profile of cardiac surgical patients using multivariable prediction models
- Establish prediction model performance



Illustrative example
EuroSCORE

Patient related factors			Cardiac related factors		
Age [1] (years)	0	0	NYHA	select	0
Gender	select	0	CCS class 4 angina [8]	no	0
Renal impairment [2] *See calculator below for creatinine clearance*	normal (CC >85ml/min)	0	LV function	select	0
Extracardiac arteriopathy [3]	no	0	Recent MI [9]	no	0
Poor mobility [4]	no	0	Pulmonary hypertension [10]	no	0
Previous cardiac surgery	no	0	**Operation related factors**		
Chronic lung disease [5]	no	0	Urgency [11]	elective	0
Active endocarditis [6]	no	0	Weight of the intervention [12]	isolated CABG	0
Critical preoperative state [7]	no	0	Surgery on thoracic aorta	no	0
Diabetes on insulin	no	0			
EuroSCORE II — *EuroSCORE II*	0				
Note: This is the 2011 EuroSCORE II	Calculate	Clear			

Step 1
Formulating the review question and protocol

UMC Utrecht

Step 1
Formulating the review question and protocol

- Describe rationale, objectives, design, methodology and statistical considerations of the systematic review
- Define the PICOTS

Extensively discussed in the CHARMS workshop!

Step 1
Formulating the review question and protocol

Predictive performance of EuroSCORE

Population	Patients undergoing coronary artery bypass grafting
Intervention	The (additive) EuroSCORE model
Comparator	Not applicable
Outcome(s)	All cause mortality
Timing	30 days, predicted using peri-operative conditions
Setting	risk stratification in the assessment of cardiac surgical results

Step 2
Formulating the search strategy

UMC Utrecht

Step 2
Formulating the search strategy

- Use information from the PICOTS
- Combine with existing search filters
- Evaluate citations of the development paper

Tools: electronic databases, conference abstracts, hand searching, online registers

Step 3
Critical appraisal

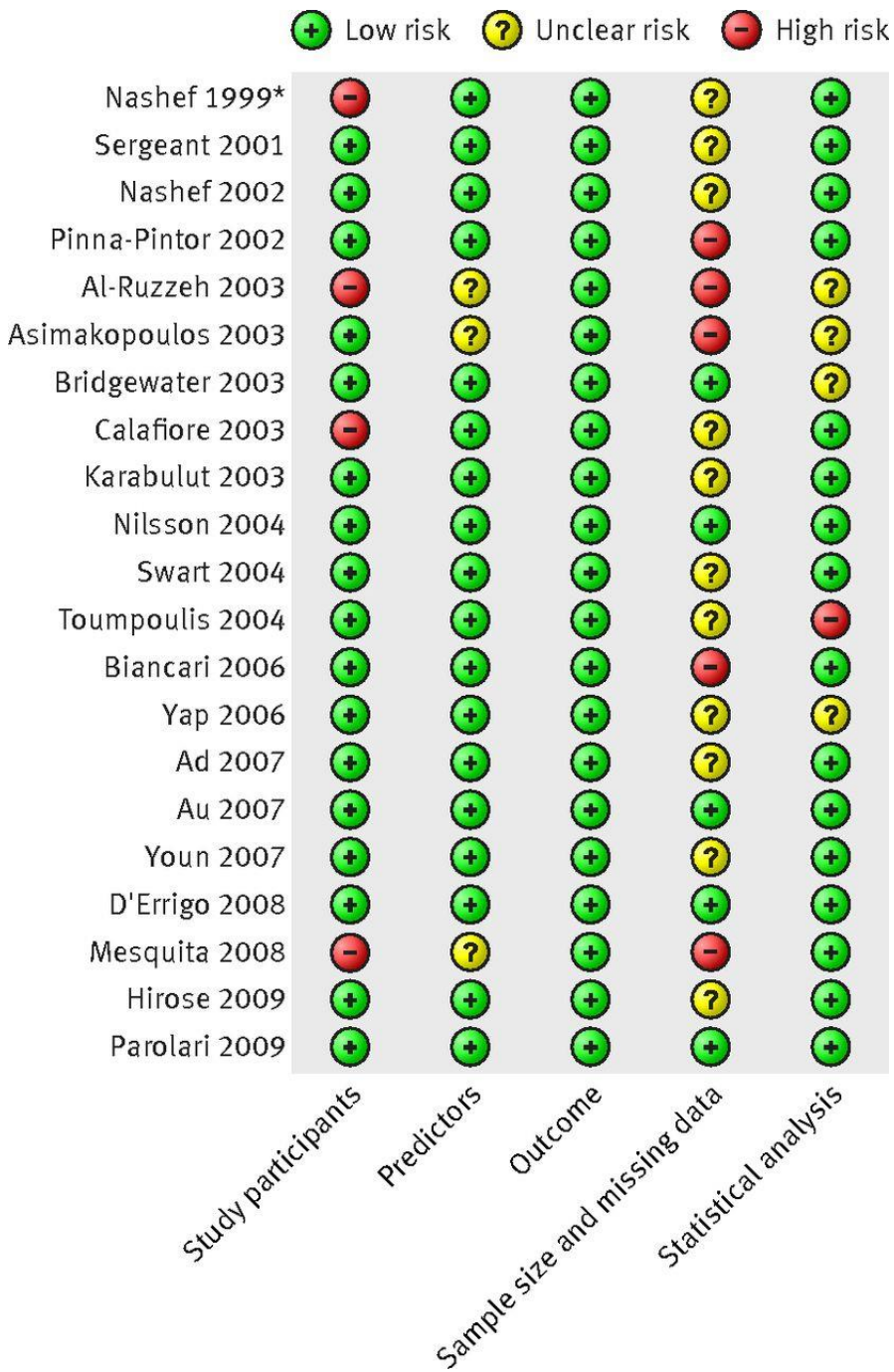
UMC Utrecht

Step 3
Critical appraisal

Evaluate **bias and applicability** of each validation study

- CHARMS checklist
- PROBAST (2017) – see previous workshop

Decide whether studies should be excluded due to low quality and/or applicability with respect to the current review



Overall judgment for risk of bias of included articles

(21 studies, involving 22 validations)

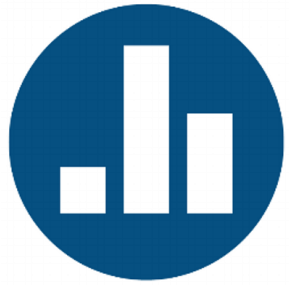
Step 4

Quantitative data extraction and preparation

UMC Utrecht

Step 4

Quantitative data extraction and preparation



What statistics can we summarize when reviewing external validation studies?

Step 4

Quantitative data extraction and preparation

What statistics can we summarize?

- Overall performance
  R^2, Brier score
- Model discrimination
  c-statistic
- Model calibration
  O:E ratio, calibration slope

Step 4

Quantitative data extraction and preparation

Common problems in data extraction

- Selective reporting
- Inconsistent measures of model performance
- Incomplete assessments (e.g. calibration)
- Missing estimates of precision (e.g. standard error)

Approximations needed to restore missing information on model performance

Step 4

Quantitative data extraction and preparation

Dealing with incomplete reporting

- C-statistic, O:E ratio and calibration slope can often be derived from reported information
- Several approximations have been proposed to obtain estimates for missing standard errors

RESEARCH METHODS AND REPORTING

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

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

Step 4

Quantitative data extraction and preparation

Software (R)

metamisc: Diagnostic and Prognostic Meta-Analysis

Meta-analysis of diagnostic and prognostic modeling studies. Summarize estimates of diagnostic test accuracy and prediction model performance. Validate, update and combine published prediction models.

Version:	0.1.5
Depends:	R (≥ 2.10), stats, graphics
Imports:	metafor, mvtnorm, ellipse, lme4
Suggests:	runjags, rjags
Published:	2017-06-22
Author:	Thomas Debray
Maintainer:	Thomas Debray <thomas.debray at gmail.com>
License:	GPL-2
URL:	http://r-forge.r-project.org/projects/metamisc/
NeedsCompilation:	no
In views:	MetaAnalysis
CRAN checks:	metamisc results

Step 4

Quantitative data extraction and preparation

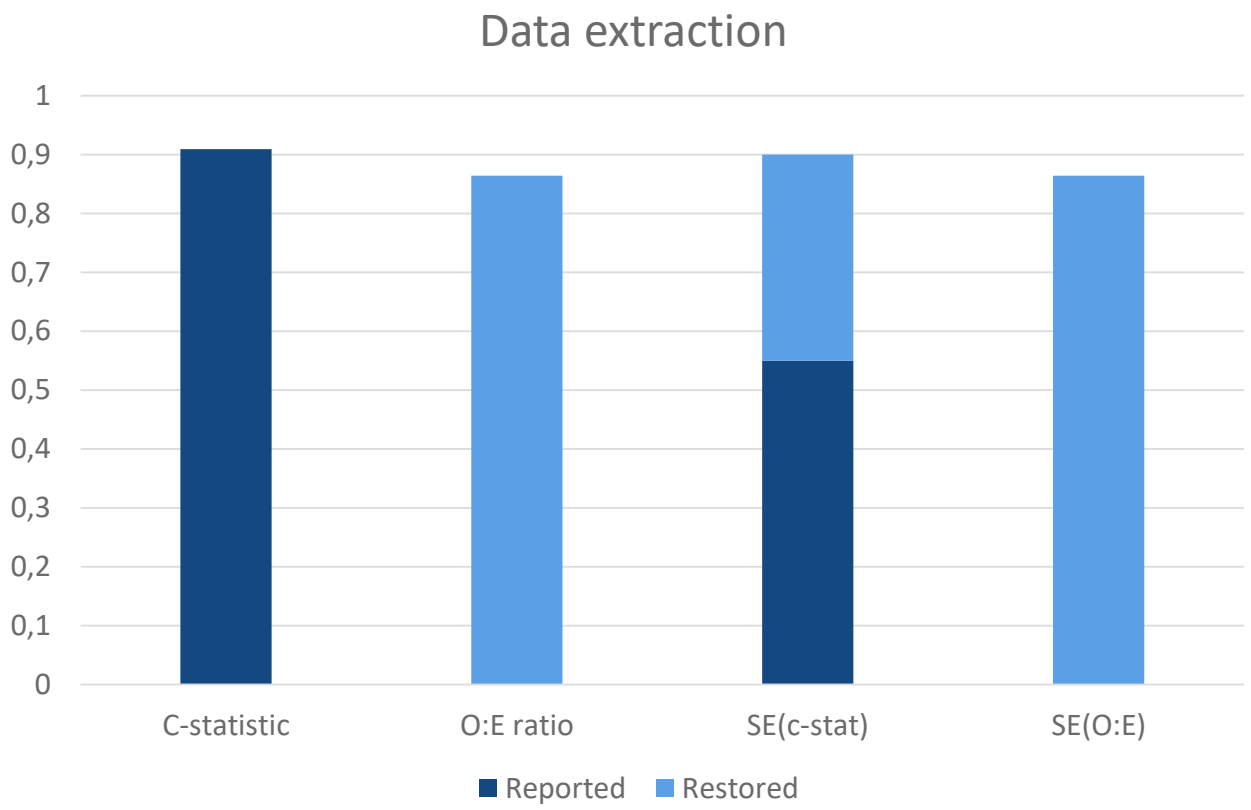
Predictive performance of the EuroSCORE

- C-statistic
  - Summary statistic reported in 20 validations
  - SE approximated for 7 studies
- O:E
  - Relevant information obtained for 21 validations
- Case-mix
  - Distribution of the LP obtained for 15 validation studies

Step 4

Quantitative data extraction and preparation

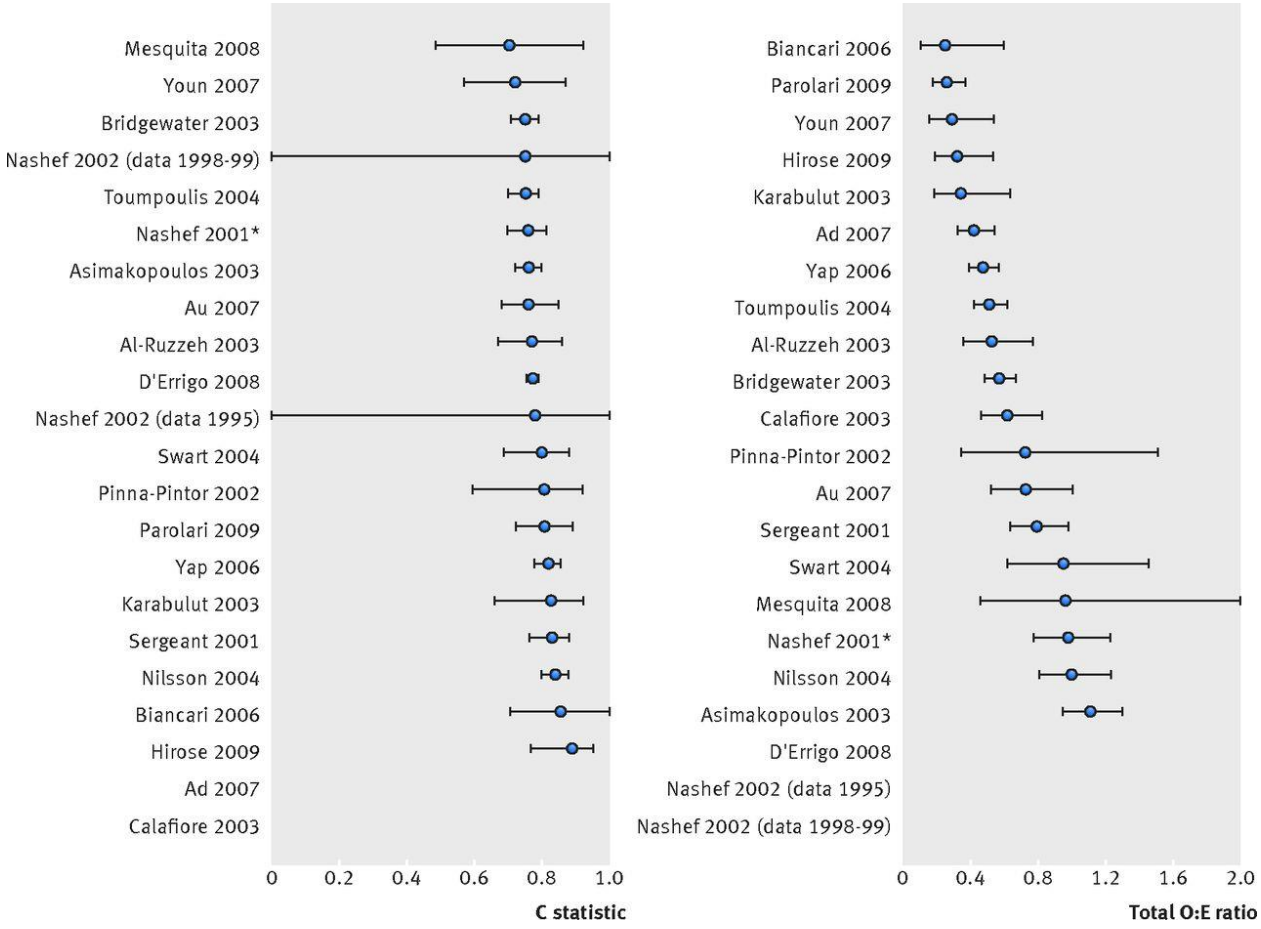
Predictive performance of the EuroSCORE



Step 4

Quantitative data extraction and preparation

Predictive performance of the EuroSCORE



Step 4

Quantitative data extraction and preparation

Other information to extract

- Information on case-mix variation
  - Mean & standard deviation of key subject characteristics
  - Mean & standard deviation of the linear predictor
- Information on key study characteristics
  - Location
  - Standards w.r.t. treatments, patient referral, …

Step 5
Meta-analysis

UMC Utrecht

Step 5
Meta-analysis

Fixed or random effects?

- Fixed effect meta-analysis
  - The model's *true* predictive accuracy is the same for all validation studies
  - Variation in predictive accuracy only appears due to chance
- Random effects meta-analysis
  - The model's *true* predictive accuracy differs across validation studies
  - Variation in predictive accuracy arises from sampling error and between-study heterogeneity

Step 5
Meta-analysis

Fixed or random effects?

- Assumption of homogeneity (fixed effect) often unrealistic because validation studies typically differ in design, execution and case-mix variation
- Ignoring heterogeneity leads to an overly precise summary result
- Summary estimates of predictive accuracy have limited usefulness when there is strong heterogeneity

Step 5
Meta-analysis

Other considerations

- Traditional meta-analysis methods assume normality of performance statistics within and across studies
- Normality assumption often challenged because:
  - Some performance measures are bounded: c-statistic (between 0 and 1), total O:E ratio (between 0 and +Inf)
  - Central Limit Theorem not applicable in small samples
- Potentially leading to misleading estimates of uncertainty, and to biased summary estimates

Step 5
Meta-analysis

Recommendations

- Allow for random effects
- Rescaling of C-statistics using **logit** transformation
- Rescaling of total O:E ratios using **log** transformation
- No rescaling needed for calibration slope or calibration-in-the-large
- Apply restricted maximum likelihood estimation
- Use Hartung-Knapp-Sidik-Jonkman method for deriving 95% confidence intervals

Step 5
Meta-analysis

Recommendations

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
0(0) 1–18

DOI: 10.1177/0962280217705678
journals.sagepub.com/home/smm
SAGE

Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures?

Kym IE Snell,[1] Joie Ensor,[1] Thomas PA Debray,[2,3] Karel GM Moons[2,3] and Richard D Riley[1]

Step 5
Meta-analysis

Software (R)

metamisc: Diagnostic and Prognostic Meta-Analysis

Meta-analysis of diagnostic and prognostic modeling studies. Summarize estimates of diagnostic test accuracy and prediction model performance. Validate, update and combine published prediction models.

Version:	0.1.5
Depends:	R (≥ 2.10), stats, graphics
Imports:	metafor, mvtnorm, ellipse, lme4
Suggests:	runjags, rjags
Published:	2017-06-22
Author:	Thomas Debray
Maintainer:	Thomas Debray <thomas.debray at gmail.com>
License:	GPL-2
URL:	http://r-forge.r-project.org/projects/metamisc/
NeedsCompilation:	no
In views:	MetaAnalysis
CRAN checks:	metamisc results

Step 5
Meta-analysis

Quantifying heterogeneity

Prediction interval

- Combines the standard error of the summary estimate with the estimate for between-study variability
- Typically based on Student's t distribution
- Provides a range for the potential predictive accuracy in a new validation study
- Ideally calculated from 10 or more validation studies

Step 5
Meta-analysis

Quantifying heterogeneity

Probability of "good" performance

- Calculate the likelihood of achieving a certain c-statistic and/or total O:E ratio in a new validation study
- Rough indication of model generalizability

Ref: Snell et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model . JCE 2015.

Step 5
Meta-analysis

Results for EuroSCORE

Meta-analysis	N	Summary	95% CI	95% PI
C-statistic	18	0.78	0.76 – 0.80	0.73 – 0.83
O:E ratio	19	0.55	0.43 – 0.69	0.20 – 1.53

Probability of "good" discrimination ($c > 0.75$) = **89%**

Probability of "good" calibration ($0.8 \leq O:E \leq 1.2$) = **15%**

Step 6

Investigating heterogeneity across studies

UMC Utrecht

Step 6

Investigating heterogeneity across studies

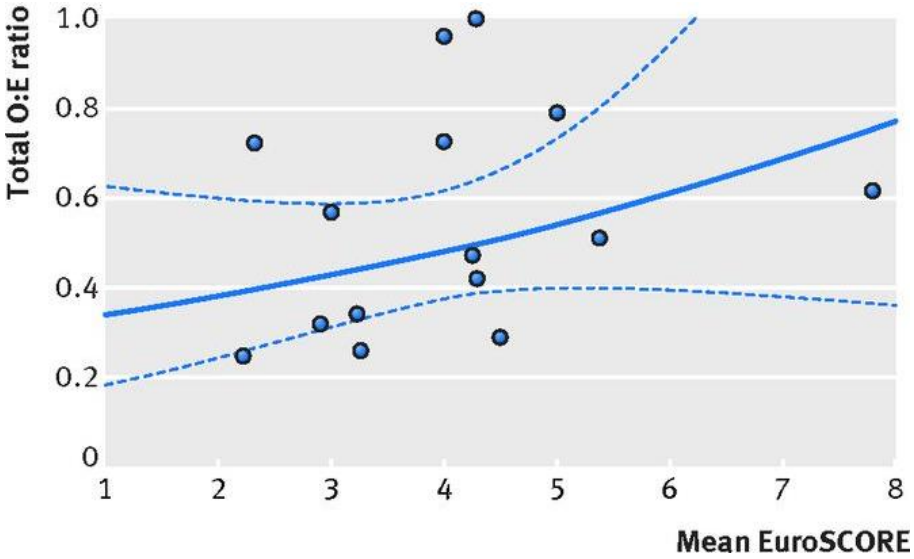
- Summary estimates of limited value in presence of strong heterogeneity
- Heterogeneity in model performance should be expected
  - C statistic may vary due to differences in “true” regression coefficients and/or due to differences in case-mix
  - Total O:E ratio may vary due to differences in outcome prevalence
- Need for meta-regression / subgroup analysis

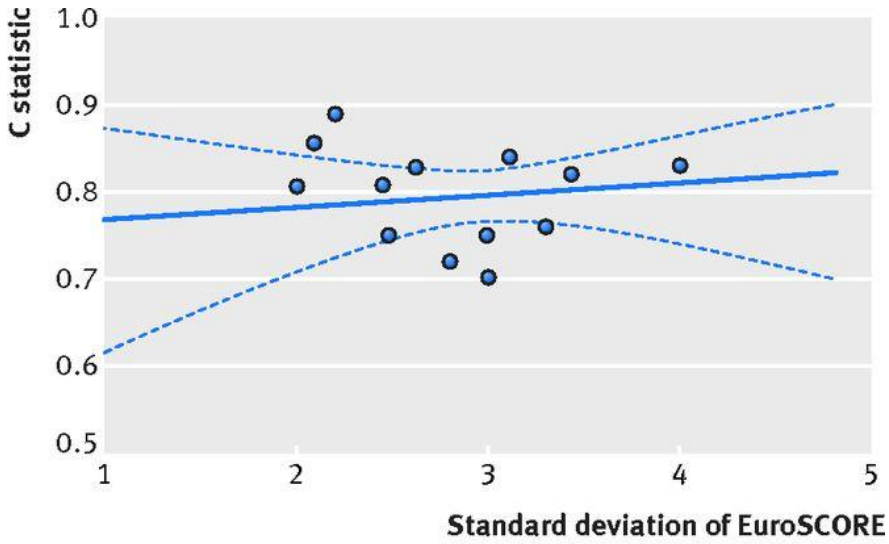
Step 6

Investigating heterogeneity across studies

Meta-analysis of EuroSCORE performance

Adjustment for case-mix variation





Step 7

Sensitivity analyses

UMC Utrecht

Step 7
Sensitivity analyses

Evaluate the robustness of drawn conclusions

- Influence of low(er) quality validation studies
- Influence of key modelling assumptions
- …

Step 7
Sensitivity analyses

Results for EuroSCORE

Meta-analysis	ROB	M	Summary	95% CI	95% PI
C-statistic	All	18	0.78	0.76 – 0.80	0.73 – 0.83
	Low	4	0.80	0.73 – 0.85	0.66 – 0.89
O:E ratio	All	19	0.55	0.43 – 0.69	0.20 – 1.53
	Low	3	0.57	0.10 – 3.33	0.02 – 19.15

Step 7
Sensitivity analyses

Multivariate meta-analysis

- Joint pooling of model discrimination and calibration
- Borrow information across different performance measures within and across studies
- Make joint inferences on different aspects of model performance in new populations

ELSEVIER

Journal of Clinical Epidemiology

Volume 69, January 2016, Pages 40–50

Original Article

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

Kym I.E. Snell[a], Harry Hua[b], Thomas P.A. Debray[c, d], Joie Ensor[e], Maxime P. Look[f], Karel G.M. Moons[c, d], Richard D. Riley[e]

Step 8
Reporting

UMC Utrecht

Step 8
Reporting

Relevant guidelines

- PRISMA
- TRIPOD
- GRADE

Case study

Performance of the Pooled Cohort Equations

UMC Utrecht

Step 1
Formulating the review question and protocol

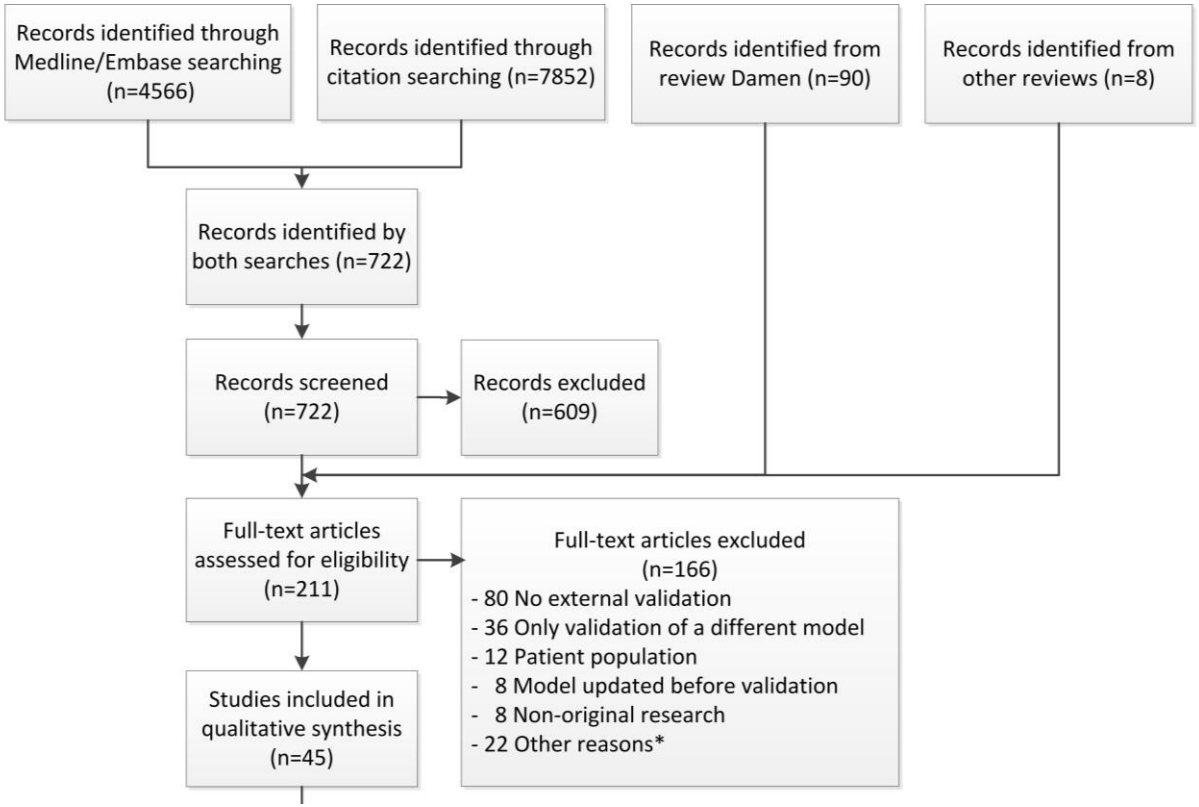
Predictive performance of PCE

Population	General population
Intervention	PCE
Comparator	Framingham Wilson and ATP III
Outcome(s)	Cardiovascular Disease (CVD)
Timing	10 year
Setting	Primary care and public health

Step 2

Formulating the search strategy

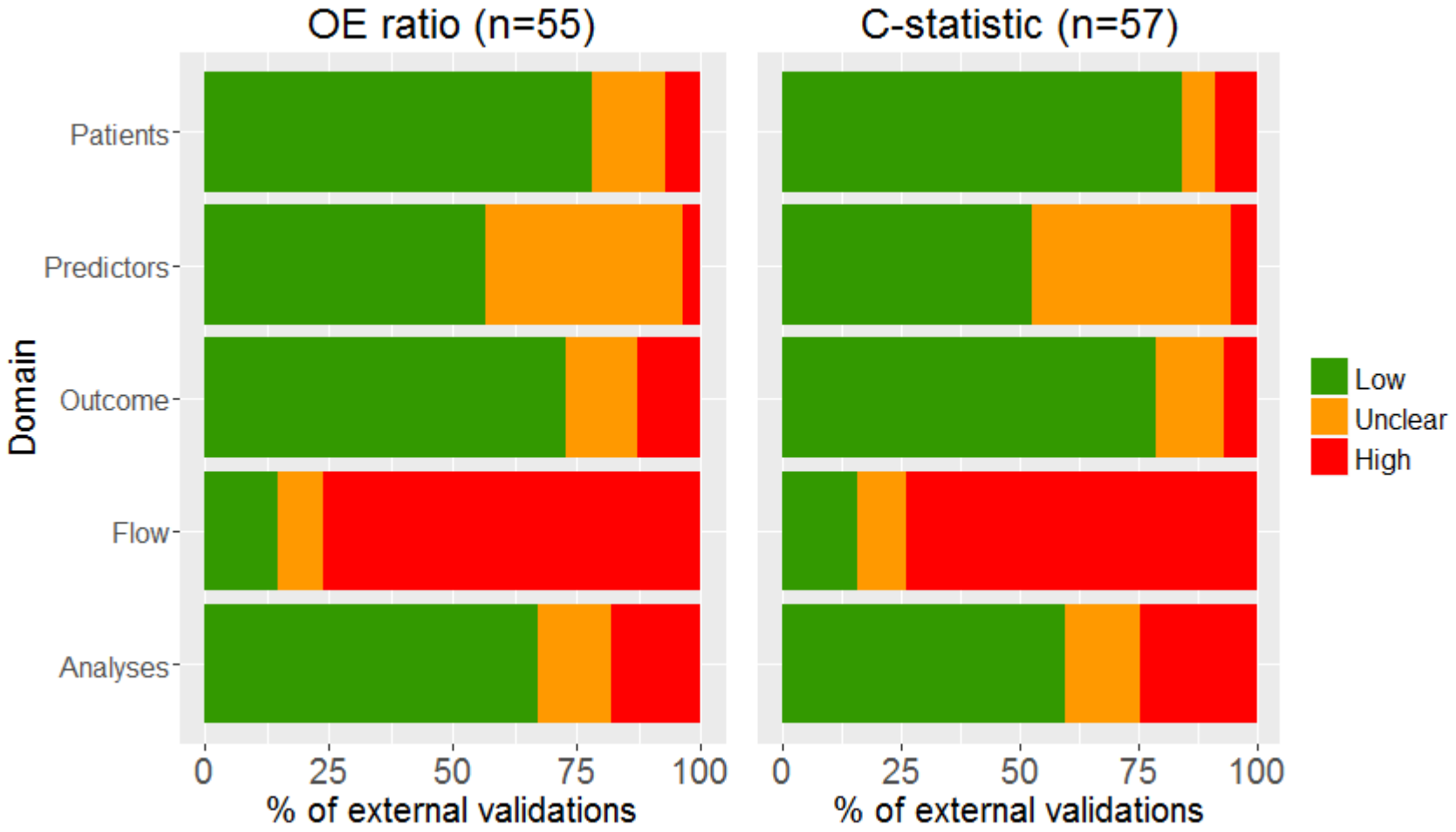
- Articles published before June 2013 selected from a previous review[1]
- Update using citation search



[1]BMJ 2016;353:i2416

Step 3
Critical appraisal

- Risk of bias assessed using a preliminary version of PROBAST



Step 4

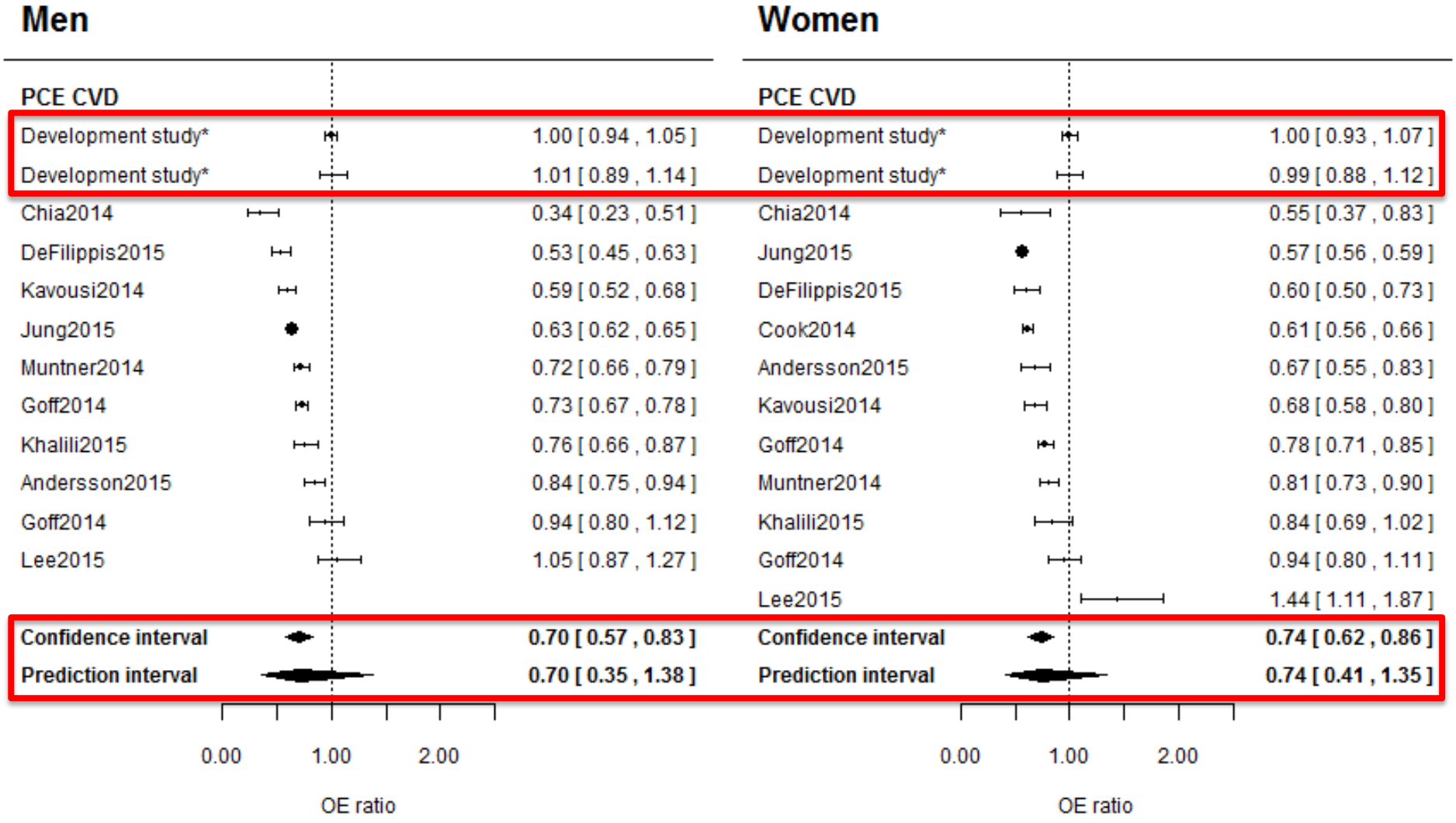
Quantitative data extraction and preparation

Items extracted:

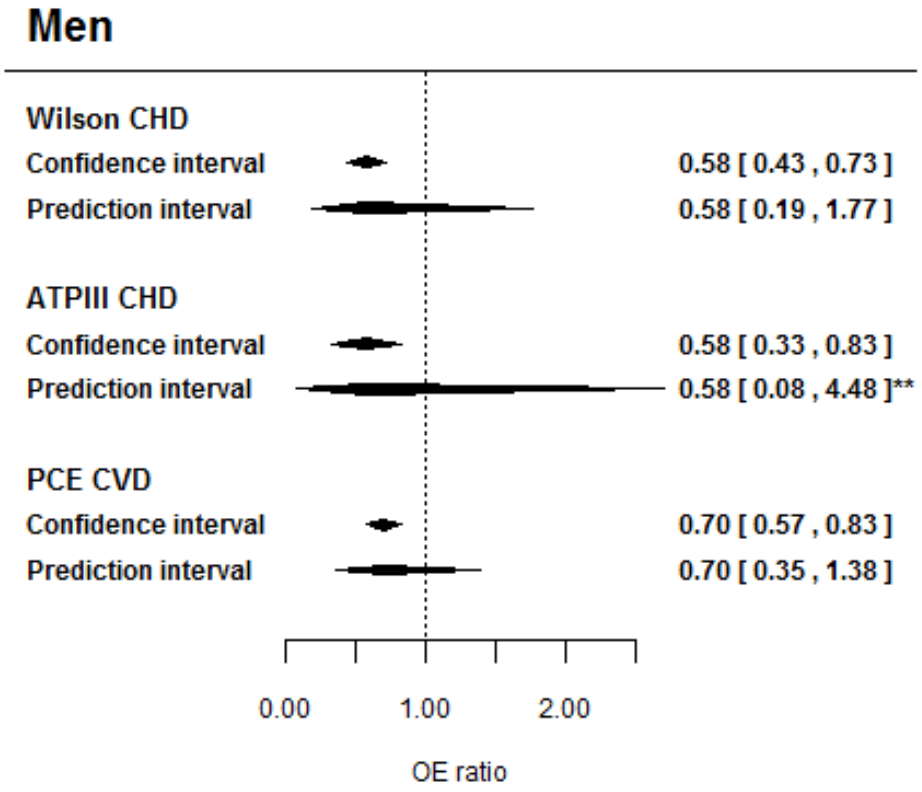
- Study design
- Study population, location
- Study dates
- Case mix
- Predictors
- Outcome definition
- Sample size
- Model performance
  - discrimination (c-statistic)
  - calibration (OE ratio, calibration slope)

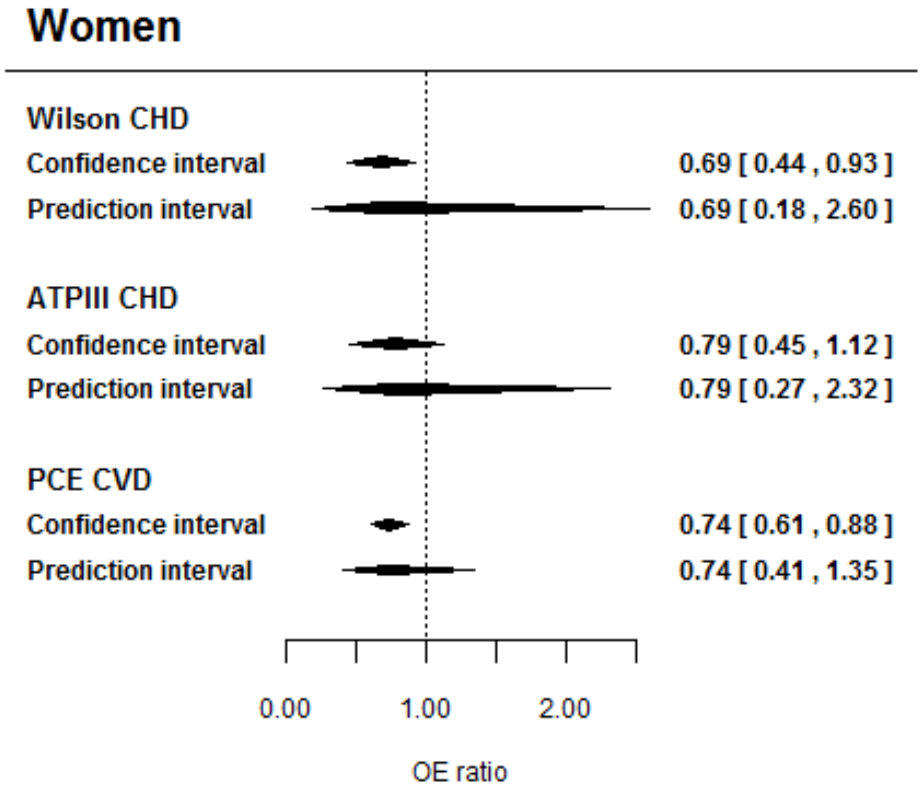
Step 5

Meta-analysis



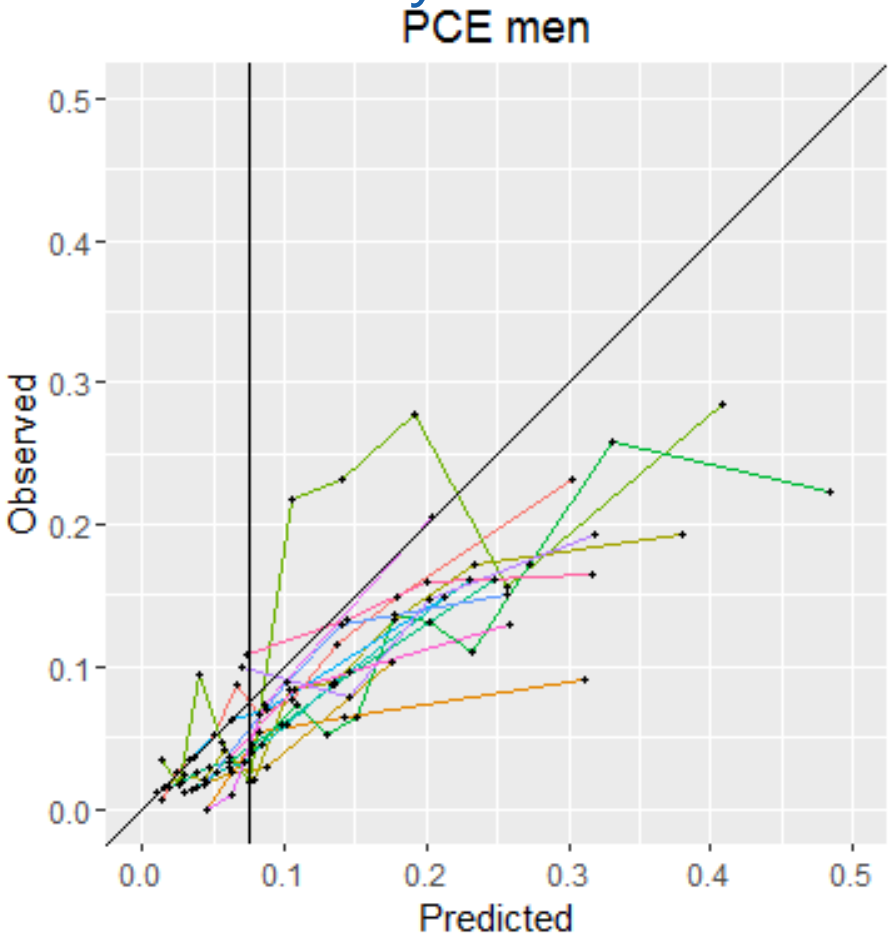
Step 5
Meta-analysis

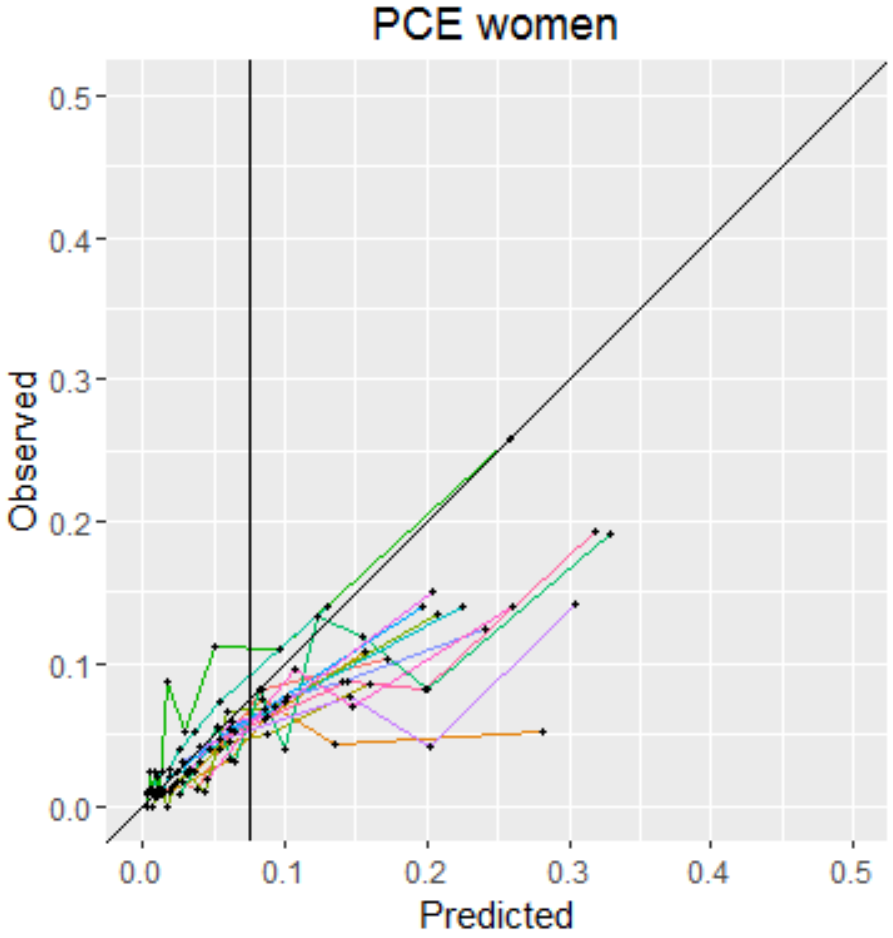




Step 5

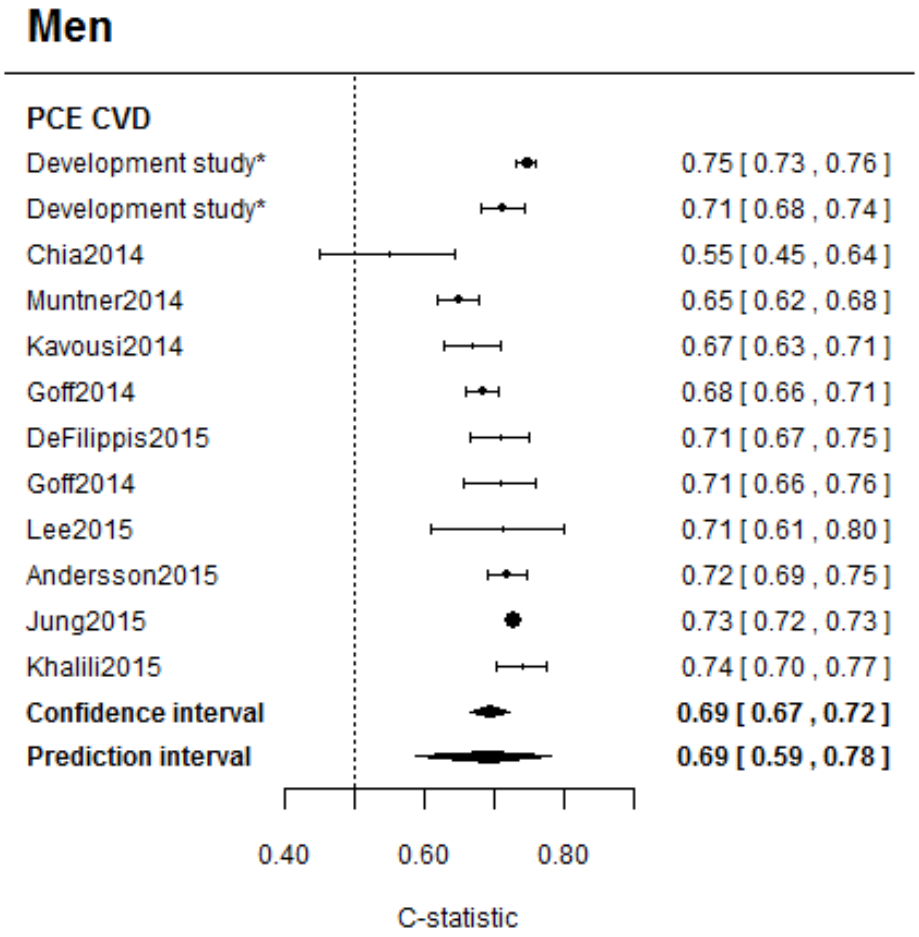
Meta-analysis

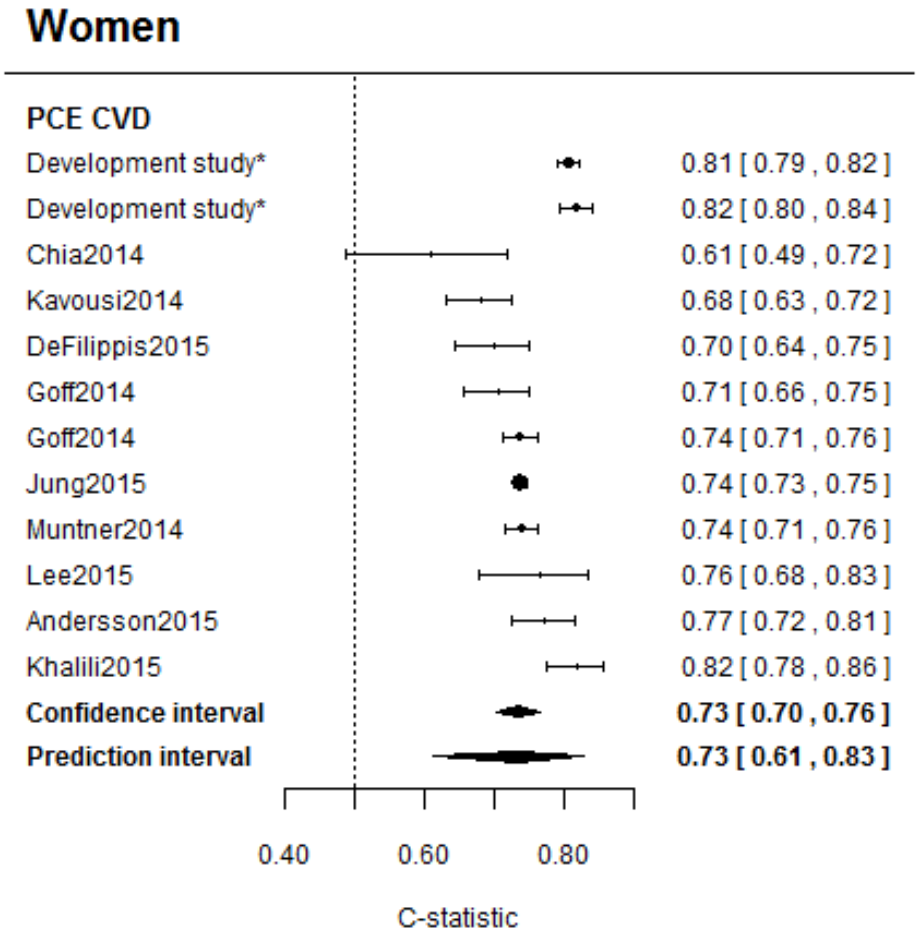




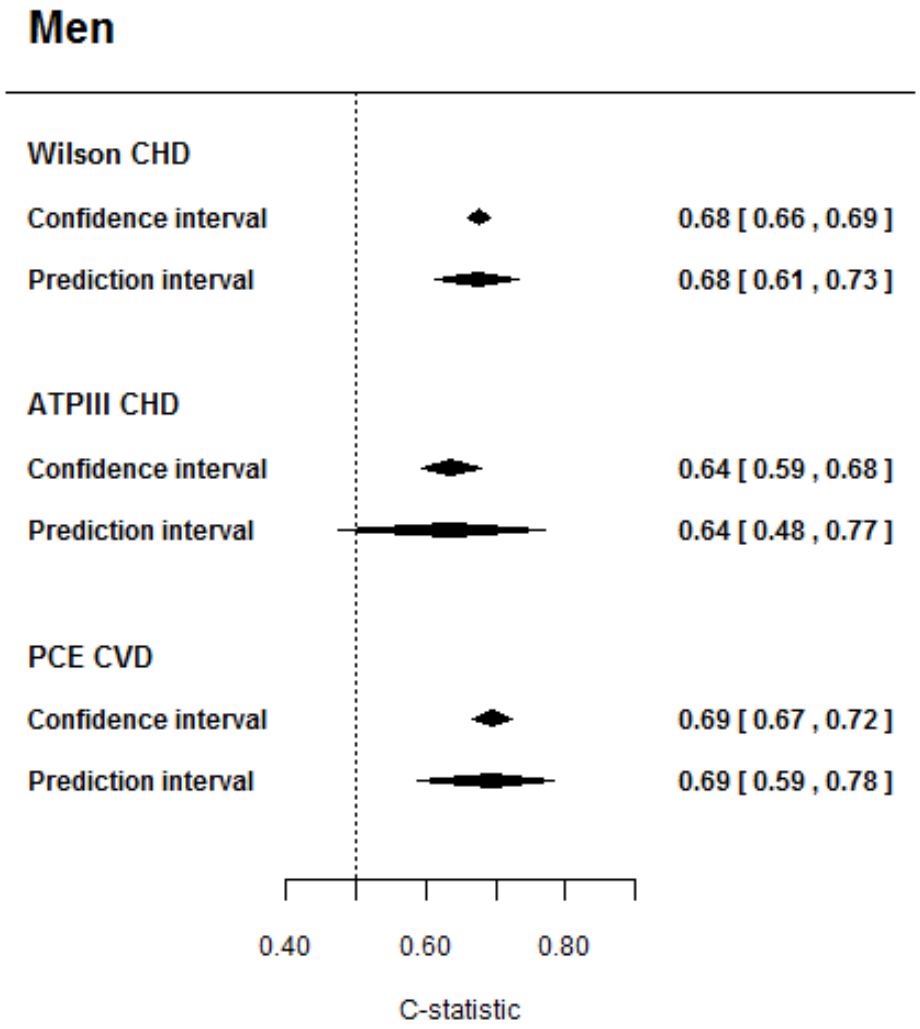
Step 5

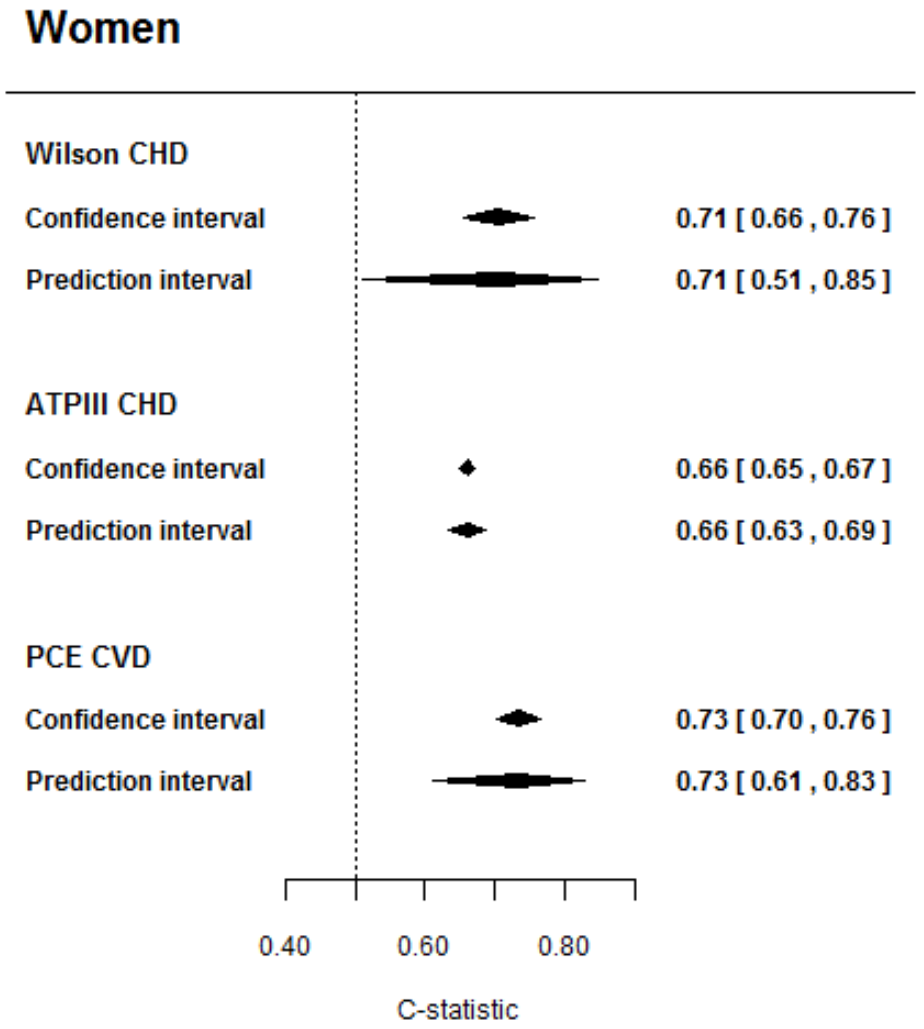
Meta-analysis





Step 5
Meta-analysis





Step 6

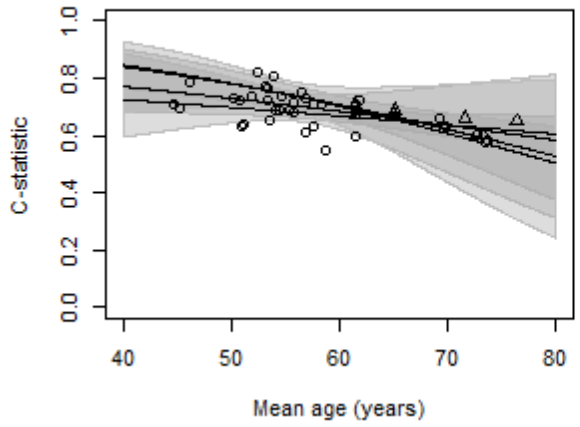
Investigating heterogeneity across studies

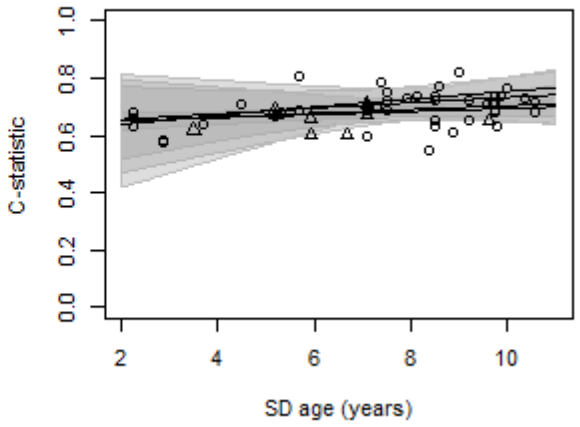
OE ratio

- Closer to 1 in US compared to other continents
- No association found for other variables (e.g. elgibility criteria, patient characteristics, year)

C-statistic

- Decrease with higher mean age, mean SBP and lower sd age
- No association found for other variables





Step 7
Sensitivity analyses

		PCE men		PCE women
OE ratio	N	OE (95%CI)	N	OE (95%CI)
All validations	10	0.698 (0.565-0.862)	11	0.742 (0.62-0.888)
Low risk of bias for all domains	2	-	3	-
Weighted by number of events	10	0.698 (0.567-0.86)	11	0.739 (0.619-0.881)
Bivariate analyses	10	0.693 (0.58-0.828)	11	0.739 (0.633-0.863)
Not extrapolated to 10 year	10	0.698 (0.565-0.862)	11	0.742 (0.62-0.888)
C-statistic	N	C (95%CI)	N	C (95%CI)
All validations	10	0.694 (0.660-0.726)	10	0.733 (0.695-0.768)
Low risk of bias for all domains	2	-	2	-
Weighted by number of events	10	0.696 (0.664-0.726)	10	0.733 (0.694-0.769)
Bivariate analyses	10	0.695 (0.665-0.724)	11	0.734 (0.703-0.762)

Closing remarks

UMC Utrecht

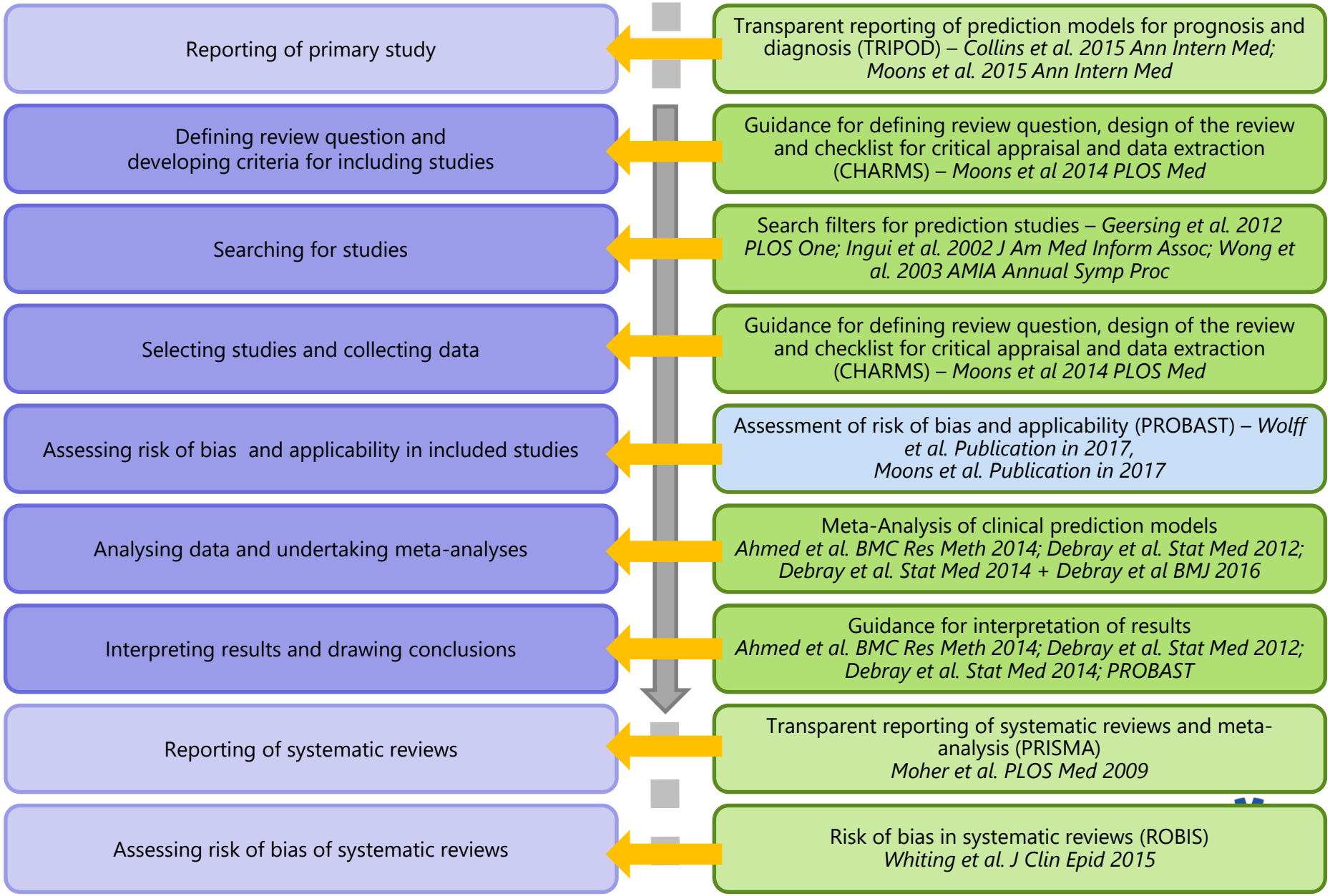
Concluding remarks

- Many similarities to other types of meta-analysis, however,
  - Data extraction more difficult
  - Heterogeneity more common
  - Summary estimates less meaningful

- Recommendations
  - Need for better reporting
  - Need for (minimal set of) standard performance measures
  - Need for IPD

- Tools for data preparation & meta-analysis
  - R package “metamisc”

Conducting systematic reviews of prediction model studies



Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 - http://handbook.cochrane.org/

Handy tools / Papers

- Debray TPA et al. **A new framework to enhance the interpretation of external validation studies of clinical prediction models**. J Clin Epidemiol. 2015.

- Debray TPA et al. **A guide to systematic review and meta-analysis of prediction model performance**. BMJ. 2017.

- Snell KIE et al. **Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model**. Journal of Clinical Epidemiology. 2015 May;69:40–50.

- Snell KIE et al. **Prediction model performance across multiple studies: which scale to use for the c-statistic and calibration measures**? Stat Met Meth Res. 2017.

Workshop aftercare

- Questions about workshop?
- Assistant needed with review of studies of prognosis studies?
- Please contact:
  - PMG Coordinator: Alexandra Hendry (Alexandra.Hendry@sswahs.nsw.gov.au)
  - PMG Co-convenor: Karel Moons (K.G.M.Moons@umcutrecht.nl)

Basic & Advanced courses

in Systematic Reviews, Meta Analysis, Clinical Epidemiolgy and Statistics

MSC EPIDEMIOLOGY

Face to Face & Online

- Systematic Reviews of Randomised Intervention Studies
- Systematic Reviews of Diagnostic Studies
- Systematic Reviews of Prognostic Studies
- Meta Analysis with Individual Participants Data
- Clinical Trials and Drug Risk Assessment
- Diagnostic Research
- Prognostic Research
- Missing Data

www.msc-epidemiology.eu

www.msc-epidemiology.online

UMC Utrecht Julius Center

Universiteit Utrecht

elevate