Systematic reviews and meta-analyses of prediction model studies: methods and applications

Anneke Damen

Systematic reviews and meta-analyses of prediction model studies: methods and applications

ISBN: 978-90-393-7005-6
Author: Anneke Damen
Cover design: Anneke Damen, Thomas van der Vlis, Lotty Hooft
Lay-out: Thomas van der Vlis, Persoonlijk Proefschrift, Utrecht, the Netherlands
Printed by: IPSKAMP printing, Enschede, the Netherlands

Studies in this thesis were funded by the Netherlands Organization for Scientific Research. Financial support by Cochrane Netherlands and the Julius Center for Health Sciences and Primary Care for the publication of this thesis is gratefully acknowledged. Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged.

Systematic reviews and meta-analyses of prediction model studies: methods and applications

Systematische literatuuroverzichten en meta-analyses van predictiemodellen: methoden en toepassingen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op donderdag 14 juni 2018 des middags te 2.30 uur

door

Johanna Antonia Adriana Gerdina Damen

geboren op 10 oktober 1990
te Oosterhout

Promotoren:	Prof.dr. K.G.M. Moons
	Prof.dr. R.J.P.M. Scholten
Copromotoren:	Dr. L. Hooft
	Dr. T.P.A. Debray

“Never doubt your heart, it’s always in its place”

K’s Choice in “Along For The Ride”

Contents

Chapter 1

General introduction

Diagnosis and prognosis are an essential part of medicine. Especially in this era of personalized medicine, predictive information about someone's diagnosis and prognosis is vital.

What are prediction models and how are they being used?

Prediction models are being used to determine an individual's risk profile. They can be used to estimate the current outcome (e.g. disease status) of an individual, in case of diagnostic models, or predict future outcome status, in case of prognostic models.[1-3] Prognostic models can estimate the future health status of currently healthy individuals, in which case they are used for targeting primary prevention (e.g. the 10-year risk of cardiovascular disease (CVD) in the general population), to estimate the chance of a recurrent event for targeting secondary prevention (e.g. the 1-year risk of stroke in patients with TIA), or to predict the course of disease in participants who already have a disease (e.g. the 3-month risk of mortality in patients with liver cirrhosis).[1,2]

Prediction models are a combination of two or more predictors that are associated with, but not necessarily causing, the outcome of interest.[4] Predictors can be patient characteristics (such as demographics, symptoms, signs, comorbidities, anthropometrics), but also results of tests (such as imaging, laboratory tests, or genetic tests), disease characteristics (such as disease stage), and others.[3] In development studies, the predictive effect of these factors are assessed. Predictors with the strongest independent predictive effect on the outcome receive the highest weight in the model.[3]

Prognostic models can be useful to make individualized decisions on preventative treatment, such as lifestyle interventions or risk lowering drugs.[2] In clinical practice, models can be used to enhance decision making on treatment administration,[5] and to assist in the communication about the course of a disease between physicians and patients.[2,6] In a research setting, prediction models are for example used to stratify patients by disease severity, or to correct for confounders in observational causal studies.[2,6]

How are prediction models being assessed?

The performance of a prediction model is often measured in terms of discrimination and calibration.[2,3,7] With discrimination we mean the ability of a model to distinguish between individuals with and without the outcome of interest. This is often quantified by the area under the receiver operating curve (AUC) or concordance (c)-statistic. An AUC or c-statistic of 1 means perfect discriminative ability, whereas a model with a c-statistic of 0.5 is not better than flipping a coin.[8] Calibration is the agreement, on average, between the number of participants with the outcome as predicted by the model versus the actually observed number of participants with the outcome. Calibration is often quantified by the OE ratio: the number of observed participants with the outcome (O) divided by the expected number of participants with the outcome (E), as predicted

by the model. Ideally, calibration is presented in plots or tables per risk category (e.g. low, medium and high categories, or in tenths of predicted risk).[2,9]

These prediction model performance measures can directly be calculated in the dataset used for model development, called apparent performance, or after applying some form of so-called internal validation of the prediction model in the development dataset.[2,3] Frequently used methods for internal validation are bootstrapping and cross-validation.[6] Internally validating a prediction model gives information about the reproducibility of the prediction model. Ideally, however, we also have information about the transportability and generalizability to other populations or settings, which is done by testing the performance in an independent dataset that was not used for development of the model.[10,11] This is called external validation. External validation studies provide information about the performance of a model in populations that, for example, differ in the case-mix from the development population or in situations where predictors or outcomes are measured and defined in different ways. All this variation between development and external validation datasets may cause a model to predict less well in external validation samples, and result in conflicting conclusions about the predictive performance of a model.[6,10,12] Changes in predictors during follow-up, e.g. due to effective treatments that will modify the occurrence of the outcome, may also affect the performance of a prediction model in the validation dataset.[13,14] Therefore, it is advised to perform multiple external validation studies to get full insight in the transportability and usefulness of a prediction model.[10]

If the performance of a prediction model is not sufficient enough when tested in a validation dataset, the model can be updated or predictors can be added based on the validation dataset at hand. Updating means that the model is tailored to the new setting, e.g. by adjusting the strength of the predictor weights (beta coefficients) or correcting for differences in outcome frequency between development and validation dataset by adjusting the intercept term of the model.[15,16] Incremental value studies assess the added value of a predictor on top of the predictors already included in a prediction model.[17] After prediction model development and validation (with or without model updating or extending) the final step in the evaluation of prediction models is to quantify to what extent the actual use of a (validated) prediction model impacts medical decision making and participant outcomes, compared to not using that prediction model. This is ideally studied in prospective studies with a comparative, randomized, design.[15,18]

The role of systematic reviews in the development and use of prediction models

Many systematic reviews have shown that studies in which prediction models are being developed, do not use recommended methods. For example, model development studies may have a very low sample size, which in turn may lead to overfitted models resulting in reduced generalizability and thus in reduced usefulness of the model.[19-21] Continuous variables are often being categorized, which may also reduce the generalizability of prediction models because not all available information is used.[19,22-24] Participants with missing data may have been excluded which may lead to selection bias.[19,23,25,26] Predictors during model development may be selected based on univariable analyses, which may result in overfitted models or missing important predictors.[19,22,27,28]

Systematic reviews have also shown that many models exist for the same target population and predicted outcome.[22,23,29-32] For healthcare professionals it can be very difficult to choose which model to use for their patients, and to what extent the predicted risks are sufficiently accurate in their own setting. Therefore, systematic reviews (with or without meta-analyses) are becoming increasingly important to overview the evidence on existing models that are developed and/or validated in a certain medical domain or setting. For systematic reviews of randomized trials ample guidance on methods, conduct, and reporting is available, but this guidance hardly exists for systematic reviews of prediction models.[33]

Objective

The aim of this thesis was to provide guidance on how to perform systematic reviews and meta-analyses of prediction model studies and to apply the developed guidance on prediction model studies in the field of cardiovascular disease (CVD).

Outline of this thesis

In **Chapter 2** we present a guide for systematic reviews and meta-analyses of prediction models.

In the field of CVD, many challenges arise. There is an overabundance of prediction models that are poorly developed, reported and validated. In **Chapter 3** we provide an overview of all existing prediction models for predicting future occurrence of CVD in the general population.

Since many competing models exist in this CVD field and external validation studies of these models often report conflicting results, in **Chapter 4** we meta-analysed the predictive performance of three frequently advocated prediction models to predict 10-year risk of CVD.

Identification of new predictors may help improve the predictive performance of existing CVD risk prediction models. In **Chapter 5** we studied the incremental value of multiple biomarkers over existing predictors to predict 10-year risk of CVD.

Use of treatments during follow-up may affect the predictive performance of a model and can be a source of heterogeneity in predictive performance of a prediction model across study populations.[13,14] **Chapter 6** gives an overview on how treatment use is currently being handled in prediction models for CVD.

Due to poor reporting of prediction model studies,[19] it is often difficult or even impossible to validate developed models or apply them in clinical practice. In order to improve the quality of reporting, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement has been published, which serves as a reporting guideline for prediction model studies.[34,35] In **Chapter 7** we present the results of a baseline measurement on the quality of reporting before the introduction of the TRIPOD statement.

Finally, we noticed that heterogeneity in reported performance of prediction models is not only a problem in the field of CVD, but also in many other clinical domains. As variations in study design and quality can partly explain this heterogeneity[2,4,10,36,37] we studied this using a meta-epidemiological approach, as presented in **Chapter 8**.

This thesis ends with a general discussion.

References

1. Grobbee DE, Hoes AW. Clinical Epidemiology - Principles, Methods and Applications for Clinical Research: London: Jones and Bartlett Publishers, 2009.
2. Steyerberg E. Clinical prediction models: a practical approach to development, validation, and updating: Springer Science & Business Media, 2008.
3. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012;98(9):683-90.
4. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375.
5. Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ 2013;346:e5793.
6. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19(4):453-73.
7. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21(1):128-38.
8. Harrell F. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis: Springer, 2015.
9. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167-76.
10. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2014.
11. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130(6):515-24.
12. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.
13. Groenwold RH, Moons KG, Pajouheshnia R, Altman DG, Collins GS, Debray TP, et al. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. J Clin Epidemiol 2016;78:90-100.
14. Pajouheshnia R, Peelen LM, Moons KGM, Reitsma JB, Groenwold RHH. Accounting for treatment use when validating a prognostic model: a simulation study. BMC Med Res Methodol 2017;17(1):103.
15. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98(9):691-8.

16. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol 2008;61(1):76-86.
17. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. Eur J Clin Invest 2012;42(2):216-28.
18. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.
19. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med 2012;9(5):1-12.
20. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995;48(12):1503-10.
21. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49(12):1373-9.
22. Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. Cancer Invest 2009;27(3):235-43.
23. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med 2011;9:103.
24. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25(1):127-41.
25. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59(10):1087-91.
26. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995;142(12):1255-64.
27. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-87.
28. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol 1996;49(8):907-16.
29. Kleinrouweler CE, Cheong-See FM, Collins GS, Kwee A, Thangaratinam S, Khan KS, et al. Prognostic models in obstetrics: available, but far from applicable. Am J Obstet Gynecol 2015.
30. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. BMC Med Inform Decis Mak 2006;6:38.
31. van Dieren S, Beulens JW, Kengne AP, Peelen LM, Rutten GE, Woodward M, et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. Heart 2012;98(5):360-9.

32. Shariat SF, Karakiewicz PI, Roehrborn CG, Kattan MW. An updated catalog of prostate cancer predictive tools. Cancer 2008;113(11):3075-99.
33. Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. BMJ 2009;339:b4184.
34. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162(1):55-63.
35. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162(1):W1-73.
36. Janssen KJ, Vergouwe Y, Donders AR, Harrell FE, Jr., Chen Q, Grobbee DE, et al. Dealing with missing predictor values when applying clinical prediction models. Clin Chem 2009;55(5):994-1001.
37. Held U, Kessels A, Garcia Aymerich J, Basagana X, Ter Riet G, Moons KG, et al. Methods for Handling Missing Variables in Risk Prediction Models. Am J Epidemiol 2016;184(7):545-51.

Chapter 2

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

Johanna AAG Damen*
Thomas PA Debray*
Kym I E Snell
Joie Ensor
Lotty Hooft
Johannes B Reitsma
Richard D Riley
Karel GM Moons

*Authors equally contributed

BMJ. 2017;356:i6460

Abstract

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

Introduction

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

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

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

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

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

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

Empirical example

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

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

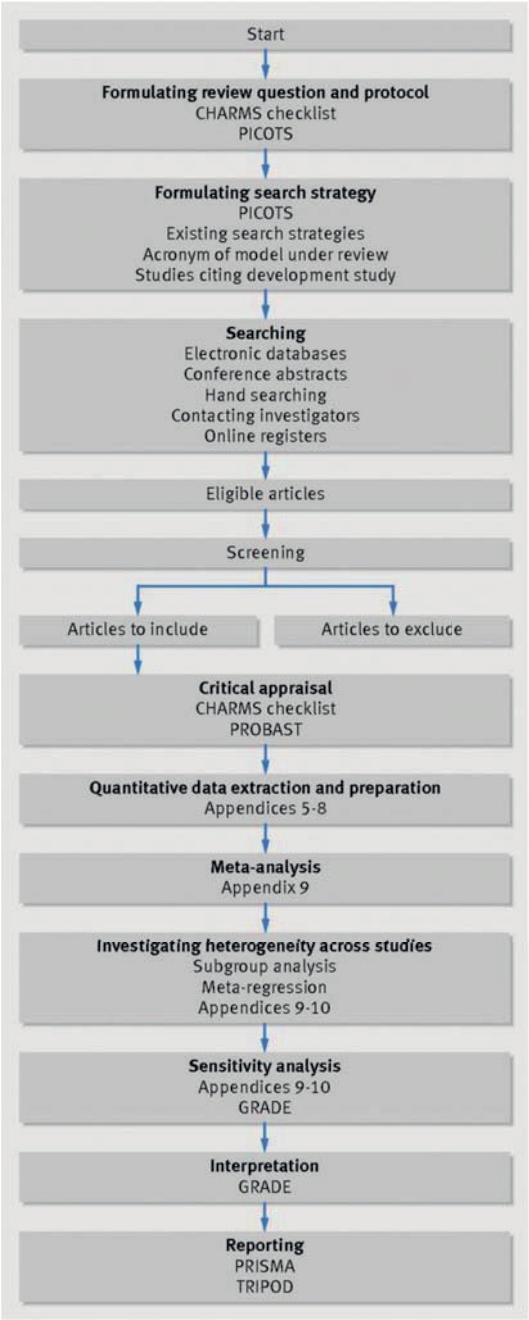


Figure 1: Flowchart for systematically reviewing and, if considered appropriate, meta-analysis of the validation studies of a prediction model. CHARMS=checklist for critical appraisal and data extraction for systematic reviews of prediction modelling studies; PROBAST=prediction model risk of bias assessment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

Steps of the systematic review

Formulating the review question and protocol

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

Case study

The formal review question was as follows: to what extent is the additive EuroSCORE able to predict all cause mortality at 30 days in patients undergoing CABG? The question is primarily interested in the predictive performance of the original EuroSCORE, and not how it performs after it has been recalibrated or adjusted in new data.

Formulating the search strategy

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

Case study

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

Box 1: PICOTS system

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

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

Critical appraisal

The quality of any meta-analysis of a systematic review strongly depends on the relevance and methodological quality of included studies. For this reason, it is important to evaluate their congruence with the review question, and to assess flaws in the design, conduct, and analysis of each validation study. This practice is also recommended by Cochrane, and can be implemented using the CHARMS checklist,[6] and, in the near future, using the prediction model risk of bias assessment tool (PROBAST).[7]

Case study

Using the CHARMS checklist and a preliminary version of the PROBAST tool, we critically appraised the risk of bias of each retrieved validation study of the EuroSCORE, as well as of the model development study. Most (n=14) of the 22 validation studies were of low or unclear risk of bias (Figure 2). Unfortunately, several validation studies did not report how missing data were handled (n=13) or performed complete case analysis (n=5). We planned a sensitivity analysis that excluded all validation studies with high risk of bias for at least one domain (n=8).[21]

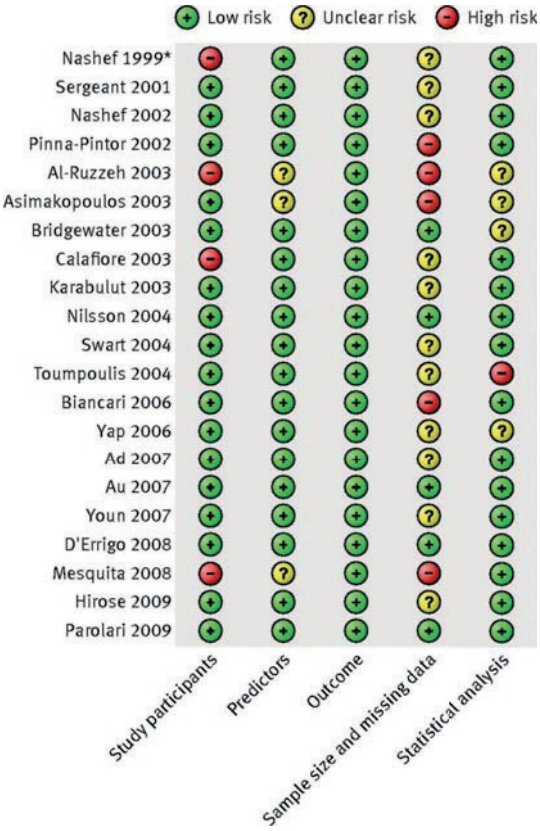


Figure 2 Overall judgment for risk of bias of included articles in the case study (predictive performance of the EuroSCORE for all cause mortality at 30 days in patients undergoing coronary artery bypass grafting). Study references listed in appendix 2. Study participants domain=design of the included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

Quantitative data extraction and preparation

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

Discrimination

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

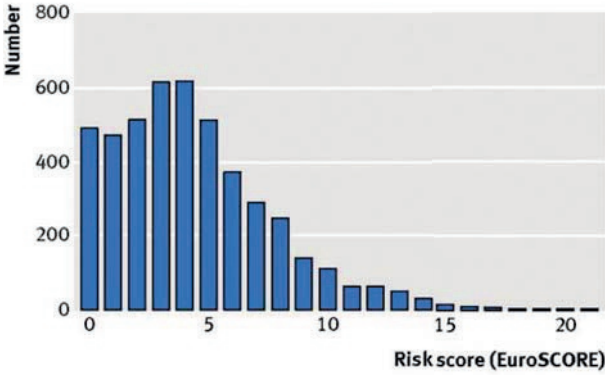
The C statistic of a prediction model can vary substantially across different validation studies. A common cause for heterogeneity in reported C statistics relates to differences between studied populations or study designs.[8,22] In particular, it has been demonstrated that the distribution of patient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

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

Case study

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

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



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

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

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

$$= 2.50$$

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

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

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

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

Study (country, enrolment year)	Validation study results							
	Total sample size	Observed deaths (total No)	Expected deaths as predicted by the model (total No)	C statistic*	EuroSCORE†		Calibration plot	Calibration table‡
					Mean	*SD*		
Nashef 1999 (8 countries, 1995)§	1497	70.6	72.4	0.7590	—	—	Absent	Present
Sergeant 2001 (Belgium, 1997-2000)	2051	81	101.8	0.83 (0.03)	5	4	Present	Present
Nashef 2002 (USA, 1995)	153 397	—	—	0.78	—	—	Absent	Absent
Nashef 2002 (USA, 1998-99)	—	—	—	0.75	—	—	Absent	Absent
Pinna-Pintor 2002 (Italy, 1993-94)	418	7	—	0.806	2.32	2.0	Present	Absent
Al-Ruzzeh 2003 (UK, 1996-2000)¶	1907	26	49.6	0.77 (0.67 to 0.86)	—	—	Absent	Present
Asimakopoulos 2003 (UK, 1993-99)¶	4654	152	137	0.76 (0.72 to 0.80)	—	—	Present	Absent
Bridgewater 2003 (UK, 1999-2002)	8572	144	257	0.75	3.0	2.48**	Present	Absent
Calafiore 2003 (Italy, 1994-2001)	1020	46	76.4	—	7.8	—	Absent	Present

Study (country, enrolment year)	Validation study results							
	Total sample size	Observed deaths (total No)	Expected deaths as predicted by the model (total No)	C statistic*	EuroSCORE†		Calibration plot	Calibration table‡
					Mean	*SD*		
Karabulut 2003 (Turkey, 1999-2001)	912	10	29.5	0.828	3.23	2.62††	Absent	Present
Nilsson 2004 (1996-2001)	4497	85	85	0.84 (0.80 to 0.88)	4.28**	3.11**	Present	Present
Swart 2004 (South Africa)	574	21	22.39	0.80	—	—	Absent	Absent
Toumpoulis 2004 (USA, 1992-2002)	3760	103	—	0.75 (0.70 to 0.79)	5.38	2.99	Absent	Present
Biancari 2006 (Finland, 1992-93)	917	5	—	0.856 (0.706 to 1.006)	2.22**	2.09**	Absent	Present
Yap 2006 (Australia, 2001-05)	5592	112	237.66	0.82	4.25	3.43††	Absent	Present
Ad 2007 (USA, 2001-04)	3125	57	134.38	—	4.3	3.2	Absent	Absent
Au 2007 (Hong Kong, 1999-2005)	1247	36	49.88	0.76 (0.68 to 0.85)	4.0	3.3	Absent	Absent
Youn 2007 (Korea, 2002-06)	757	10	34.2	0.72 (0.57 to 0.87)	4.5	2.8	Absent	Present
D'Errigo 2008 (Italy, 2002-04)	30 610	777	—	0.773 (0.755 to 0.791)	—	—	Present	Absent

Study (country, enrolment year)	Validation study results							
	Total sample size	Observed deaths (total No)	Expected deaths as predicted by the model (total No)	C statistic*	EuroSCORE†		Calibration plot	Calibration table‡
					Mean	*SD*		
Mesquita 2008 (Brazil, 2005-07)	144	7	7.34	0.702 (0.485 to 0.919)	4	3	Absent	Absent
Hirose 2009 (Japan, 1991-2006)	1522	14	—	0.890	2.9	2.2	Present	Present
Parolari 2009 (Italy, 1999-2007)	3440	29	108.88	0.808 (0.723 to 0.892)	3.26	2.45	Absent	Absent

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

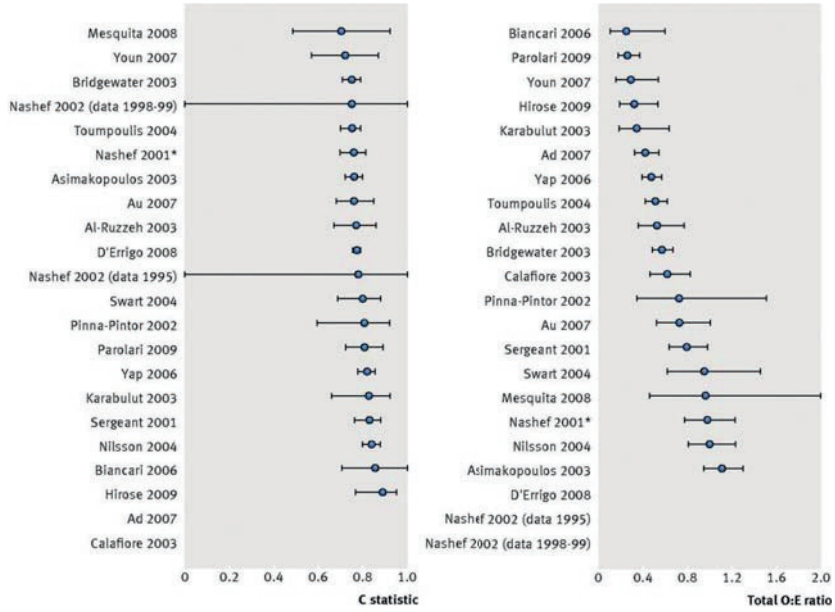


Figure 4: Forest plots of extracted performance statistics of the additive EuroSCORE in the case study (to predict all cause mortality at 30 days in patients undergoing coronary artery bypass grafting). Part A shows forest plot of study specific C statistics (all 95% confidence intervals estimated on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

Calibration

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

Summarising estimates of calibration performance is challenging because calibration plots are most often not presented, and because studies tend to report different types of summary statistics in calibration.[12,27] Therefore, we propose to extract information on the total number of observed (O) and expected (E) events, which are statistics most likely to be reported or derivable (appendix 7). The total O:E ratio provides a rough indication

of the overall model calibration (across the entire range of predicted risks). The total O:E ratio is strongly related to the calibration in the large (appendix 5), but that is rarely reported. The O:E ratio might also be available in subgroups, for example, defined by tenths of predicted risk or by particular groups of interest (eg, ethnic groups, or regions). These O:E ratios could also be extracted, although it is unlikely that all studies will report the same subgroups. Finally, it would be helpful to also extract and summarise estimates of the calibration slope.

Case study

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

Performance of survival models

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

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

Meta-analysis

Once all relevant studies have been identified and corresponding results have been extracted, the retrieved estimates of model discrimination and calibration can be summarised into a weighted average. Because validation studies typically differ in design, execution, and thus case-mix, variation between their results are unlikely to occur

by chance only.[8,22] For this reason, the meta-analysis should usually allow for (rather than ignore) the presence of heterogeneity and aim to produce a summary result (with its 95% confidence interval) that quantifies the average performance across studies. This can be achieved by implementing a random (rather than a fixed) effects meta-analysis model (appendix 9). The meta-analysis then also yields an estimate of the between-study standard deviation, which directly quantifies the extent of heterogeneity across studies.[19] Other meta-analysis models have also been proposed, such as by Pennells and colleagues, who suggest weighting by the number of events in each study because this is the principal determinant of study precision.[31] However, we recommend to use traditional random effects models where the weights are based on the within-study error variance. Although it is common to summarise estimates of model discrimination and calibration separately, they can also jointly be synthesised using multivariate meta-analysis.[9] This might help to increase precision of summary estimates, and to avoid exclusion of studies for which relevant estimates are missing (eg, discrimination is reported but not calibration).

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

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

Case study

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

Investigating heterogeneity across studies

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

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

Key issues that could be considered as modifiers of model performance are differences in the heterogeneity between patients across the included validation studies (difference case mix variation),[8] differences in study characteristics (eg, in terms of design, follow-up time, or outcome definition), and differences in the statistical analysis or characteristics related to selective reporting and publication (eg, risk of bias, study

size). The regression coefficient obtained from a meta-regression analysis describes how the dependent variable (here, the logit C statistic or log O:E ratio) changes between subgroups of studies in case of a categorical explanatory variable or with one unit increase in a continuous explanatory variable. The statistical significance measure of the regression coefficient is a test of whether there is a (linear) relation between the model's performance and the explanatory variable. However, unless the number of studies is reasonably large (>10), the power to detect a genuine association with these tests will usually be low. In addition, it is well known that meta-regression and subgroup analysis are prone to ecological bias when investigating summarised patient level covariates as modifiers of model performance.[40]

Case study

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

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

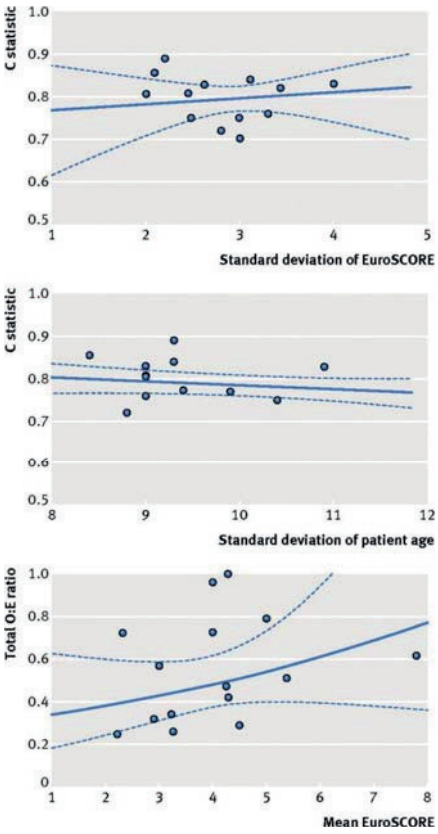


Figure 5: Results from random effects meta-regression models in the case study (predictive performance of the EuroSCORE for all cause mortality at 30 days in patients undergoing coronary artery bypass grafting). Solid lines=regression lines; dashed lines=95% confidence intervals; dots=included validation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

Sensitivity analysis

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

Case study

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

Reporting and presentation

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

As illustrated in this article, researchers should clearly describe the review question, search strategy, tools used for critical appraisal and risk of bias assessment, quality of the included studies, methods used for data extraction and meta-analysis, data used for meta-analysis, and corresponding results and their uncertainty. Furthermore, we recommend to report details on the relevant study populations (eg, using the mean and standard deviation of the linear predictor) and to present summary estimates with confidence intervals and, if appropriate, prediction intervals. Finally, it might be helpful to report probabilities of good performance separately for each performance measure, because researchers can then decide which criteria are most relevant for their situation.

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

Meta-analysis	Performance	Risk of bias	No of included studies	Summary estimate	95% confidence interval	95% prediction interval
Univariate*	C statistic	Low/ unclear/ high	18	0.78	0.76 to 0.80	0.73 to 0.83
Univariate*	O:E ratio	Low/ unclear/ high	19	0.55	0.43 to 0.69	0.20 to 1.53
Bivariate*	C statistic	Low/ unclear/ high	20	0.79	0.77 to 0.80	0.73 to 0.83
Bivariate*	O:E ratio	Low/ unclear/ high	20	0.55	0.44 to 0.68	0.20 to 1.47

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

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

Concluding remarks

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

statistics, quantitatively summarise the extracted estimates, and investigate sources of between-study heterogeneity.

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

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

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

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

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

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

Acknowledgements

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

References

1. Khan K, Kunz R, Kleijnen J, Antes G. *Systematic reviews to support evidence-based medicine*: Crc Press, 2011.
2. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10(2):e1001381.
3. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. *PLoS One* 2012;7(2):e32844.
4. Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. *AMIA Annu Symp Proc* 2003:728-32.
5. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc* 2001;8(4):391-7.
6. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11(10):e1001744.
7. Wolff R, Collins GS, Kleijnen J, Mallett S, Reitsma JB, Riley R, et al. PROBAST: a risk of bias tool for prediction modelling studies. *24th Cochrane Colloquium*. Seoul, South Korea: Cochrane Database of Systematic Reviews, 2016.
8. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2014.
9. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2015.
10. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19(4):453-73.
11. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130(6):515-24.
12. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013;66(3):268-77.
13. Siregar S, Groenwold RH, de Heer F, Bots ML, van der Graaf Y, van Herwerden LA. Performance of the original EuroSCORE. *Eur J Cardiothorac Surg* 2012;41(4):746-54.
14. Echouffo-Tcheugui JB, Batty GD, Kivimaki M, Kengne AP. Risk models to predict hypertension: a systematic review. *PLoS One* 2013;8(7):e67370.
15. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* 2009;302(21):2345-52.

16. Eichler K, Puhan MA, Steurer J, Bachmann LM. Prediction of first coronary events with the Framingham score: a systematic review. *Am Heart J* 2007;153(5):722-31, 31.e1-8.
17. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak* 2006;6:38.
18. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;14:40.
19. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;353:i3140.
20. Peat G, Riley RD, Croft P, Morley KI, Kyzas PA, Moons KG, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. *PLoS Med* 2014;11(7):e1001671.
21. Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. *BMJ* 2015;350:h870.
22. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010;172(8):971-80.
23. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.
24. van Klaveren D, Gonen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings. *Stat Med* 2016;35(23):4136-52.
25. Higgins JPT, Green S. Combining Groups.
26. http://handbook.cochrane.org/chapter_7/7_7_3_8_combining_groups.htm, 2011.
27. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. *BMC Med Res Methodol* 2005;5:13.
28. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9(5):1-12.
29. Austin PC, Pencinca MJ, Steyerberg EW. Predictive accuracy of novel risk factors and markers: A simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model. *Stat Methods Med Res* 2017;26(3):1053-77.
30. Blanche P, Dartigues JF, Jacqmin-Gadda H. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biom J* 2013;55(5):687-704.

31. Jinks RC, Royston P, Parmar MK. Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. *BMC Med Res Methodol* 2015;15:82.
32. Pennells L, Kaptoge S, White IR, Thompson SG, Wood AM. Assessing risk prediction models using individual participant data from multiple studies. *Am J Epidemiol* 2014;179(5):621-32.
33. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011;342:d549.
34. Snell KIE. Development and application of statistical methods for prognosis research. University of Birmingham, 2015.
35. van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. *BMC Med Res Methodol* 2014;14:5.
36. Gengsheng Q, Hotilovac L. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Stat Methods Med Res* 2008;17(2):207-21.
37. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* 2014;14:25.
38. Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, et al. Random-effects meta-analysis of inconsistent effects: a time for change. *Ann Intern Med* 2014;160(4):267-70.
39. Ban JW, Emparanza JI, Urreta I, Burls A. Design Characteristics Influence Performance of Clinical Prediction Rules in Validation: A Meta-Epidemiological Study. *PLoS One* 2016;11(1):e0145779.
40. Deeks JJ, Higgins J, Altman DG. Analysing data and undertaking meta-analyses. *Cochrane handbook for systematic reviews of interventions: Cochrane book series* 2008:243-96.
41. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Stat Med* 2002;21(3):371-87.
42. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;339:b2700.
43. Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G, et al. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. *JAMA* 2015;313(16):1657-65.
44. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162(1):55-63.

45. Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS Med* 2015;12(10):e1001886.
46. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res* 2001;10(4):277-303.
47. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016;352:i6.

Supplemental material

1 The additive EuroSCORE

The composition of the additive EuroSCORE system (i.e. the risk factors, their definitions and the weights allocated to them) is given below.[10] The system is additive: to calculate the predicted risk for a patient, the scores for existing risk factors are added to give an approximate percentage predicted mortality figure.

	Definition	Score
Patient-related factors		
Age	Per 5 years or part thereof over 60 years	1
Sex	Female	1
Chronic pulmonary disease	Longterm use of bronchodilators or steroids for lung disease	1
Extracardiac arteriopathy	Any one or more of the following: claudication, carotid occlusion or > 50% stenosis, previous or planned intervention on the abdominal aorta, limb arteries or carotids	2
Neurological dysfunction	Disease severely affecting ambulation or day-to-day functioning	2
Previous cardiac surgery	Requiring opening of the pericardium	3
Serum creatinine	> 200 µmol/l preoperatively	2
Active endocarditis	Patient still under antibiotic treatment for endocarditis at the time of surgery	3
Critical preoperative state	Any one or more of the following: ventricular tachycardia or fibrillation or aborted sudden death, preoperative cardiac massage, preoperative ventilation before arrival in the anaesthetic room, preoperative inotropic support, intraaortic balloon counterpulsation or preoperative acute renal failure (anuria or oliguria < 10 ml/h)	3
Cardiac-related factors		
Unstable angina	Rest angina requiring i.v. nitrates until arrival in the anaesthetic room	2
LV dysfunction	Moderate or LVEF 30 – 50 % Poor or LVEF < 30	1 3
Recent myocardial infarct	(< 90 days)	2
Pulmonary hypertension	Systolic PA pressure > 60 mmHg	2

	Definition	Score
Operation-related factors		
Emergency	Carried out on referral before the beginning of the next working day	2
Other than isolated CABG	Major cardiac procedure other than or in addition to CABG	2
Surgery on thoracic aorta	For disorder of ascending, arch or descending aorta	3
Postinfarct septal rupture		4

LVEF = left ventricular ejection fraction; PA = pulmonary artery, CABG = coronary artery bypass grafting.

2 Validation studies in the empirical example

Below is an overview of the 21 articles that were included in our meta-analysis:

R-1: Ad N, Barnett SD, Speir AM. The performance of the EuroSCORE and the Society of Thoracic Surgeons mortality risk score: the gender factor. Interact Cardiovasc Thorac Surg. 2007 Apr;6(2):192–5.

R-2: Al-Ruzzeh S, Asimakopoulos G, Ambler G, Omar R, Hasan R, Fabri B, et al. Validation of four different risk stratification systems in patients undergoing off-pump coronary artery bypass surgery: a UK multicentre analysis of 2223 patients. Heart. 2003 Apr;89(4):432–5.

R-3: Asimakopoulos G, Al-Ruzzeh S, Ambler G, Omar RZ, Punjabi P, Amrani M, et al. An eval- uation of existing risk stratification models as a tool for comparison of surgical performances for coronary artery bypass grafting between institutions. Eur J Cardiothorac Surg. 2003 Jun;23(6):93541; discussion 941–2.

R-4: Au WK, Sun MP, Lam KT, Cheng LC, Chiu SW, Das SR. Mortality prediction in adult cardiac surgery patients: comparison of two risk stratification models. Hong Kong Med J. 2007 Aug;13(4):293–7.

R-5: Biancari F, Kangasniemi O-P, Luukkonen J, Vuorisalo S, Satta J, Pokela R, et al. EuroSCORE predicts immediate and late outcome after coronary artery bypass surgery. Ann Thorac Surg. 2006 Jul;82(1):57–61.

R-6: Bridgewater B, Grayson AD, Jackson M, Brooks N, Grotte GJ, Keenan DJM, et al. Surgeon specific mortality in adult cardiac surgery: comparison between crude and risk stratified data. BMJ. 2003 Jul 5;327(7405):13–7.

R-7: Calafiore AM, Di Mauro M, Canosa C, Di Giammarco G, Iaco AL, Contini M. Early and late outcome of myocardial revascularization with and without cardiopulmonary bypass in high risk patients (EuroSCORE > or = 6). Eur J Cardiothorac Surg. 2003 Mar;23(3):360–7.

R-8: D'Errigo P, Seccareccia F, Rosato S, Manno V, Badoni G, Fusco D, et al. Comparison be- tween an empirically derived model and the EuroSCORE system in the evaluation of hospital performance: the example of the Italian CABG Outcome Project. Eur J Cardiothorac Surg. 2008 Mar;33(3):325–33.

R-9: Hirose H, Inaba H, Noguchi C, Tambara K, Yamamoto T, Yamasaki M, et al. EuroSCORE predicts postoperative mortality, certain morbidities, and recovery time. Interact Cardiovasc Thorac Surg. 2009 Oct;9(4):613–7.

R-10: Karabulut H, Toraman F, Alhan C, Camur G, Evrenkaya S, Dadelen S, et al. EuroSCORE overestimates the cardiac operative risk. Cardiovasc Surg. 2003 Aug;11(4):295–8.

R-11: Mesquita ET, Ribeiro A, Arajo MP de, Campos LA de A, Fernandes MA, Colafranceschi AS, et al. Indicators of healthcare quality in isolated coronary artery bypass graft surgery performed at a tertiary cardiology center. Arq Bras Cardiol. 2008 May;90(5):320–3.

R-12: Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European System for Cardiac Operative Risk Evaluation (EuroSCORE). European Journal of Cardio-Thoracic Surgery: Official Journal of the European Association for Cardio-Thoracic Surgery. 1999 Jul;16(1):9–13. ***[Original development study with a split-sample validation.]***

R-13: Nashef SAM, Roques F, Hammill BG, Peterson ED, Michel P, Grover FL, et al. Validation of European System for Cardiac Operative Risk Evaluation (EuroSCORE) in North American cardiac surgery. Eur J Cardiothorac Surg. 2002 Jul;22(1):101–5. ***[Two validation studies were available from this article: one using data from 1995 and one using data from 1998-1999.]***

R-14: Nilsson J, Algotsson L, Hglund P, Lhrs C, Brandt J. Early mortality in coronary bypass surgery: the EuroSCORE versus The Society of Thoracic Surgeons risk algorithm. Ann Thorac Surg. 2004 Apr;77(4):12359; discussion 1239–40.

R-15: Parolari A, Pesce LL, Trezzi M, Loardi C, Kassem S, Brambillasca C, et al. Performance of EuroSCORE in CABG and off-pump coronary artery bypass grafting: single institution experience and meta-analysis. Eur Heart J. 2009 Feb;30(3):297–304.

R-16: Pinna-Pintor P, Bobbio M, Colangelo S, Veglia F, Giammaria M, Cuni D, et al. Inaccuracy of four coronary surgery risk-adjusted models to predict mortality in individual patients. Eur J Cardiothorac Surg. 2002 Feb;21(2):199–204.

R-17: Sergeant P, de Worm E, Meyns B. Single centre, single domain validation of the EuroSCORE on a consecutive sample of primary and repeat CABG. Eur J Cardiothorac Surg. 2001 Dec;20(6):1176–82.

R-18: Swart MJ, Joubert G. The EuroSCORE does well for a single surgeon outside Europe. Eur J Cardiothorac Surg. 2004 Jan;25(1):145–6; author reply 146.

R-19: Toumpoulis IK, Anagnostopoulos CE, DeRose JJ, Swistel DG. European system for cardiac operative risk evaluation predicts long-term survival in patients with coronary artery bypass grafting. Eur J Cardiothorac Surg. 2004 Jan;25(1):51–8.

R-20: Yap C-H, Reid C, Yii M, Rowland MA, Mohajeri M, Skillington PD, et al. Validation of the EuroSCORE model in Australia. Eur J Cardiothorac Surg. 2006 Apr;29(4):441–6; discussion 446.

R-21: Youn Y-N, Kwak Y-L, Yoo K-J. Can the EuroSCORE predict the early and mid-term mortality after off-pump coronary artery bypass grafting? Ann Thorac Surg. 2007 Jun;83(6):2111–7.

3 The PICOTS system

The PICOTS system as presented in the CHARMS checklist[9] describes key items for framing the review aim, search strategy, and study inclusion and exclusion criteria. In brief, and applied to our case study:

- **Population** - Define the target population in which the prediction model will be used. In our case study, the population of interest comprises patients undergoing CABG.
- **Intervention (Model)** - Define the prediction model(s) under review. In the case study, the focus is on the prognostic additive EuroSCORE model. Accordingly, one includes the studies that externally validated the EuroSCORE model. The question remains whether or not to include the results of the model development study as well. Including the original development study may help to understand variation in results in and between external validation studies, as we illustrate below.
- **Comparator** - If applicable, one may address competing models for the prognostic model under review. Ideally, studies are included that have compared (validated) the competing models in a head-to-head fashion, i.e. both models are applied and validated in the same subjects. The existence of alternative models was not considered in our case study, and therefore not further addressed here.
- **Outcomes** - Define the outcome(s) of interest for which the model is validated. Although the majority of validation studies use the same outcome and the same outcome definition as the original development study, a prediction model can also be validated on its ability for predicting a more or less different outcome. For example, the Framingham score was designed to estimate the 10-year risk of coronary heart disease (CHD), but has also been used to predict all cause mortality and cardiovascular disease mortality.[5,15] In our case study, the outcome was defined as all cause mortality. Papers validating the EuroSCORE model to predict other outcomes such as cardiovascular mortality were excluded.
- **Timing** - Specifically for prognostic models it is important to define when (at which point in time, so-called prognostic T0) and over what time period the outcome is predicted. Alike for different outcomes, researchers may validate the same model to predict the outcome over different time periods. For example, the Framingham score for predicting 10 year CHD risk has also been validated for 5 year and lifetime predictions.[5] We here focus on 30-day all cause mortality, predicted using preoperative conditions.
- **Setting** - Define the intended role or setting of the prognostic model. For instance, in the case study the intended use of the EuroSCORE model was to perform risk stratification in the assessment of cardiac surgical results, such that operative mortality could be used as a valid measure of quality of care.

4 Calibration of a prediction model

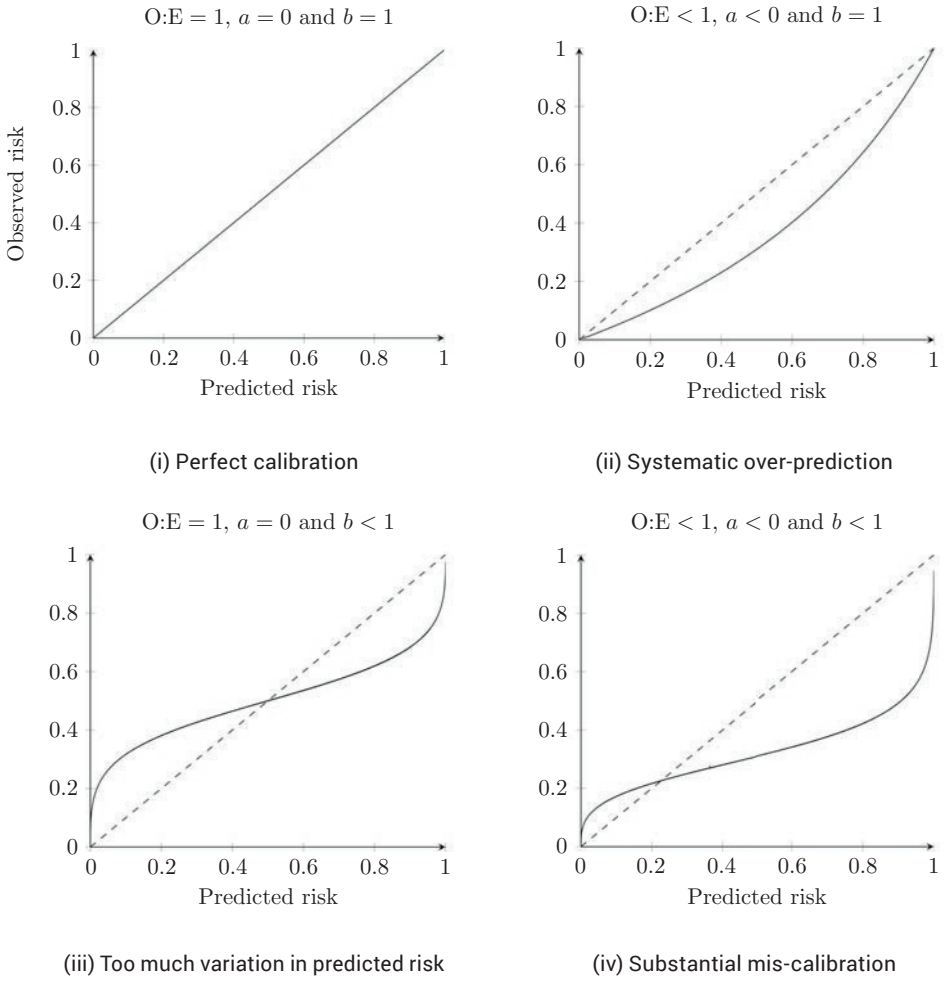


Figure 1: Calibration of a prediction model
O:E = ratio of observed versus expected risk, α = calibration-in-the-large (calculated on the logit scale), β = calibration slope (calculated on the logit scale).

Calibration plot (ii) typically occurs when the outcome occurrence in the validation set is lower than in the original development set. Calibration plot (iii) typically occurs when a prediction model was over-fitted to the development data set. Finally, calibration plot (iv) typically occurs when the prediction model was over-fitted to the development set and when the outcome occurrence in the validation set is lower than in the original development set.

5 Relation between total O:E ratio and calibration-in-the-large

For logistic regression models, calibration-in-the large (a) is calculated as follows:

$$logit(P_O) = LP + a$$

where LP indicates the average linear predictor in the validation study (using the regression coefficients of the development study), P_O indicates the observed outcome probability in the validation study and P_E indicates the expected outcome probability in the validation study. Note that for logistic regression models, we have $P_E = logit^{-1}(LP)$. Hence, when a and P_O can be extracted from a publication, we have $LP = logit(P_O) - a$.

$$O:E = \frac{O}{E} = \frac{P_O}{P_E} \quad \Leftrightarrow \quad (1)$$

$$= \frac{P_O}{logit^{-1}(LP)} \quad \Leftrightarrow \quad (2)$$

$$= \frac{P_O}{logit^{-1}(logit(P_O) - a)} \quad \Leftrightarrow \quad (3)$$

$$= \frac{P_O}{\frac{-P_O}{\exp(a)P_O - \exp(a) - P_O}} \quad \Leftrightarrow \quad (4)$$

$$= -(\exp(a)P_O - \exp(a) - P_O) \quad (5)$$

We can then use the Delta method to estimate the error variance of the total O:E ratio:[6]

$$Var(O:E) = Var(-(\exp(a)P_O - \exp(a) - P_O)) \quad \Leftrightarrow \quad (6)$$

$$= \mathrm{Var}(\exp(a)P_O) + \mathrm{Var}(\exp(a)) \quad \Leftrightarrow \quad (7)$$

$$= \left((P_O)^2 + 1\right) Var(\exp(a)) \quad \Leftrightarrow \quad (8)$$

$$\approx \left((P_O)^2 + 1\right)\left(\frac{\partial \exp(a)}{\partial a}\right)^2 Var(a) \quad \Leftrightarrow \quad (9)$$

$$\approx \left((P_O)^2 + 1\right)(\exp(a))^2 Var(a) \quad (10)$$

such that

$$Var(\ln(O:E)) \approx \left(\frac{\partial \ln(O:E)}{\partial O:E}\right)^2 Var(O:E) \quad \Leftrightarrow \quad (11)$$

$$\approx \left(\frac{\partial \ln\left(-(\exp(a)P_o - \exp(a) - P_o)\right)}{\partial a}\right)^2 \left((P_o)^2+1\right)(\exp(a))^2 Var(a) \quad \Leftrightarrow \quad (12)$$

$$\approx \left(\frac{(P_o-1)\exp(a)}{(P_o-1)\exp(a)-P_o}\right)^2 \left((P_o)^2+1\right)(\exp(a))^2 Var(a) \quad \Leftrightarrow \quad (13)$$

$$\approx \frac{(P_o-1)^2\left((P_o)^2+1\right)(\exp(P_o+a))^2}{\left(P_o(-\exp(a))+P_o+\exp(a)\right)^2} Var(a) \quad (14)$$

6 Variance estimators

Below, we provide equations for approximating the within-study error variance of ln(O:E) and logit(*c*), which is needed for meta-analysis. Let *O* denote the total number of observed events, *E* the total number of expected (predicted) events and *N* the total sample size.

In some situations, the total O:E ratio is given together with its error variance $Var(O:E)$. We can then use the Delta method to estimate $Var(\ln(O:E))$:[6]

$$Var(\ln(O:E)) \approx \left(\frac{\partial \ln(O:E)}{\partial (O:E)}\right)^2 Var(O:E) \quad \Leftrightarrow \quad (15)$$

$$\approx \frac{1}{(O:E)^2} Var(O:E) \quad (16)$$

In most situations, however, *O* and *E* are reported separately without any estimate of uncertainty. In the following derivations, we regard E as a fixed constant. We treat *O* as a binomially distributed variable since *O* is given as the number of successes (events) from *N* subjects:

$$Var(O:E) = Var\left(\frac{O}{E}\right) \quad \Leftrightarrow \quad (17)$$

$$= \frac{1}{E^2} Var(O) \quad \Leftrightarrow \quad (18)$$

$$= \frac{1}{E^2} N(P_O(1-P_O)) \quad \Leftrightarrow \quad (19)$$

$$= \frac{1}{E^2} O(1-P_O) \quad (20)$$

such that

$$Var(\ln(O:E)) = Var\left(\ln\left(\frac{O}{E}\right)\right) \quad \Leftrightarrow \quad (21)$$

$$= Var(\ln(O) - \ln(E)) \quad \Leftrightarrow \quad (22)$$

$$= Var(\ln(O)) \quad \Leftrightarrow \quad (23)$$

$$\approx \left(\frac{\partial \ln(O)}{\partial O}\right)^2 Var(O) \quad \Leftrightarrow \quad (24)$$

$$\approx \frac{1}{O^2} Var(O) \quad \Leftrightarrow \quad (25)$$

$$\approx \frac{1}{O^2} N P_O (1-P_O) \quad \Leftrightarrow \quad (26)$$

$$\approx \frac{1-P_O}{O} \quad (27)$$

For those situations in which *N* is large and P_O is very small, the Poisson distribution can be used to approximate the binomial distribution such that:

$$Var(O:E) = Var\left(\frac{O}{E}\right) \quad \Leftrightarrow \quad (28)$$

$$\approx \frac{1}{E^2} Var(O) \quad \Leftrightarrow \quad (29)$$

$$\approx \frac{O}{E^2} \quad (30)$$

Again, we can use the Delta method to estimate the within-study variance of the logarithm of the total O:E ratio:[6]

$$Var(\ln(O:E)) = Var\left(\ln\left(\frac{O}{E}\right)\right) \quad \Leftrightarrow \quad (31)$$

$$= Var(\ln(O)-\ln(E)) \quad \Leftrightarrow \quad (32)$$

$$= Var(\ln(O)) \quad \Leftrightarrow \quad (33)$$

$$\approx \left(\frac{\partial \ln(O)}{\partial O}\right)^2 Var(O) \quad \Leftrightarrow \quad (34)$$

$$\approx \left(\frac{1}{O^2}\right) O \quad \Leftrightarrow \quad (35)$$

$$\approx \frac{1}{O} \quad (36)$$

In some situations, articles report E:O = E/O = $1/O:E$ with corresponding estimates of uncertainty. We can again use the Delta method to obtain $Var(\ln(O:E))$:

$$Var(\ln(O:E)) \approx \left(\frac{\partial \ln(O:E)}{\partial (O:E)}\right)^2 Var(O:E) \quad \Leftrightarrow \quad (37)$$

$$\approx \left(\frac{1}{(O:E)^2}\right) Var(O:E) \quad \Leftrightarrow \quad (38)$$

$$\approx \frac{E^2}{O^2} Var\left(\frac{1}{E:O}\right) \quad \Leftrightarrow \quad (39)$$

$$\approx \frac{E^2}{O^2}\left(\frac{\partial (1/E:O)}{\partial (E:O)}\right)^2 Var(E:O) \quad \Leftrightarrow \quad (40)$$

$$\approx \frac{E^2}{O^2}\left(\frac{1}{(E:O)^2}\right)^2 Var(E:O) \quad \Leftrightarrow \quad (41)$$

$$\approx \frac{E^2}{O^2}\frac{O^4}{E^4} Var(E:O) \quad \Leftrightarrow \quad (42)$$

$$\approx \frac{O^2}{E^2} Var(E:O) \quad \Leftrightarrow \quad (43)$$

$$\approx (O:E)^2 Var(E:O) \quad (44)$$

Note that for prediction models with a survival (time-to-event) outcome, the number of observed events *O* is usually affected by censoring and therefore not reliable for estimating the total O:E ratio at a certain time point. We therefore propose to derive (or approximate) the observed event risk P_O from Kaplan-Meier estimates or Kaplan-Meier curves. The total O:E ratio is then given as P_O/P_E, with an error variance of:

$$Var(O:E) = Var\left(\frac{P_O}{P_E}\right) \quad \Leftrightarrow \quad (45)$$

$$= \frac{1}{P_E^2} Var(P_O) \quad (46)$$

When applying the log transformation, we have:

$$Var(\ln(O:E)) = Var\left(\ln\left(\frac{P_O}{P_E}\right)\right) \quad \Leftrightarrow \quad (47)$$

$$= Var(\ln(P_O) - \ln(P_E)) \quad \Leftrightarrow \quad (48)$$

$$= Var(\ln(P_O)) \quad \Leftrightarrow \quad (49)$$

$$\approx \left(\frac{\partial \ln(P_O)}{\partial P_O}\right)^2 Var(P_O) \quad \Leftrightarrow \quad (50)$$

$$\approx \frac{1}{P_O^2} Var(P_O) \quad (51)$$

For model discrimination, we can use the Delta method to approximate the within-study variance of the logit c-statistic:

$$Var(logit(c)) \approx \left(\frac{\partial logit(c)}{\partial c}\right)^2 Var(c) \quad \Leftrightarrow \quad (52)$$

$$\approx \left(\frac{1}{c}+\frac{1}{1-c}\right)^2 Var(c) \quad \Leftrightarrow \quad (53)$$

$$\approx \left(\frac{1}{c(1-c)}\right)^2 Var(c) \quad \Leftrightarrow \quad (54)$$

$$\approx \frac{Var(c)}{(c(1-c))^2} \quad (55)$$

When the within-study variance of the c-statistic, $\mathrm{Var}(c)$, is not known it is still possible to approximate the within-study variance of the logit c-statistic:[7,11]

$$Var(logit(c)) \approx \left(\frac{\partial logit(c)}{\partial c}\right)^2 Var(c) \quad \Leftrightarrow \quad (56)$$

$$\approx \frac{Var(c)}{(c(1-c))^2} \quad \Leftrightarrow \quad (57)$$

$$\approx \frac{(c(1-c)[1+s^*(1-c)/(2-c)+t^* c/(1+c)])/st}{(c(1-c))^2} \quad \Leftrightarrow \quad (58)$$

$$\approx \frac{[1+s^*(1-c)/(2-c)+t^* c/(1+c)]}{stc(1-c)} \quad (59)$$

where s is the total number of observed events (also denoted as O in this article), t is the total number of non-events (which can be calculated as $N - O$) and $s^* = t^* = \frac{1}{2}(s+t) - 1$.

7 Data extraction

In this section we describe how to obtain estimates for logit(*c*) and the total O:E ratio, as well as their corresponding standard error, when they are not reported. Let *O* denote the total number of observed events, *E* the total number of expected (predicted) events and *N* the total sample size. Further, we define $P_O = O/N$ as the observed, and $P_E = E/N$ as the expected event probability.

When pooling estimates of a model's discrimination and calibration, standard errors of both quantities need to be retrieved from each report. These can directly be obtained in a meta-analysis from the reported upper and lower limit of the confidence interval, or from the reported p-value.[1,2] When no appropriate estimates of uncertainty are reported, it is still possible to approximate the standard error from the total number of observed events, the total number of expected events and the total sample size.[7,11] Details are provided in Appendix 5, in Appendix 6 and in the tables below.

Example

In the study Sergeant 2001, we have *N* = 2051, *O* = 81 and *E* = 101.8 such that:

$$\begin{aligned} \ln(O:E) &= \ln(81) - \ln(101.8) \\ &= -0.23 \end{aligned}$$

and

$$\begin{aligned} SE(\ln(O:E)) &= \frac{\sqrt{(2051 \times (81/2051) \times (1 - 81/2051))}}{81} \\ &= 0.11 \end{aligned}$$

Alternatively, using Poisson approximation, we have:

$$\begin{aligned} SE(\ln(O:E)) &= \frac{1}{\sqrt{81}} \\ &= 0.11 \end{aligned}$$

Table A.1: Formulas for estimating the (logit) c-statistic and its variance from other information in a primary study

What is reported?	Estimate for $logit(c)$	Estimate for $Var(logit(c))$	Reference
c-statistic or AUC	$logit(c)$	$\left[(logit(c_{ub})-logit(c_{lb}))/(2\times 1.96)\right]^2$	[2]
		$\frac{Var(c)}{(c(1-c))^2}=\left(\frac{SE(c)}{c(1-c)}\right)^2$	[Appendix 6]
		$\frac{1+(N/2-1)(1-c)/(2-c)+(N/2-1)c/(1+c)}{c(1-c)O(N-O)}$	[Appendix 6]
Somer's D statistic (D_{yx})	$logit((D_{yx}+1)/2)$	$SE(D_{yx})/(2c(1-c))$	[12]
Distribution of the LP	$logit\left(\phi\left(\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)\right)$	No formula yet published, use one of the approximations above	[4]
Log-odds ratio for the LP (calibration slope b)	$logit\left(\phi(\sigma b/\sqrt{2})\right)$	No formula yet published, use one of the approximations above	[4]
Cohen's effect size (d)	$logit\left(\phi(d/\sqrt{2})\right)$	No formula yet published, use one of the approximations above	[4]

With $logit(c)=\ln(c)-\ln(1-c)$. The lower and upper bound of the 95% confidence interval of the c-statistic are given by c_{lb} and, respectively, by c_{ub}.
AUC: the Area Under the receiver operating characteristic Curve.

$\phi()$: the standard normal cumulative distribution function.
LP: The linear predictor (LP) is the linear combination of the model predictors in the validation study weighted by the regression coefficients of the model in the development study.

σ^2: Common variance of the LP

μ_A: Mean of the LP in the affected (Y = 1) population

σ_A^2: Variance of the LP in the affected (Y = 1) population

μ_B: Mean of the LP in the unaffected (Y = 0) population

σ_B^2: Variance of the LP in the unaffected (Y = 0) population

Table A.2: Formulas for estimating the (log) O:E ratio from other information in a primary study

What is reported?	Estimate for $\ln(O:E)$	Estimate for $Var(\ln(O:E))$	Reference
O:E	$\ln(O:E)$	$\left[\left(\ln(O:E_{ub})-\ln(O:E_{lb})\right)/(2\times 1.96)\right]^2$	[2]
		$\left[\frac{\ln(O:E)}{-0.862+\sqrt{0.743-2.404\ln(p)}}\right]^2$	[1]
O and E	$\ln(O)-\ln(E)$	$\frac{1-P_O}{O}\approx\frac{1}{O}$	[Appendix 6]
P_O, P_E and $Var(P_O)$	$\ln(P_O)-\ln(P_E)$	$\frac{1}{P_O^2}Var(P_O)$	[Appendix 6]
O and E across different risk strata	$\ln\left(\sum_i O_i\right)-\ln\left(\sum_i E_i\right)$ or $\ln\left(\sum_j P_{Oi}N_i\right)-\ln\left(\sum_j P_{Ei}N_i\right)$	$\sum_i Var(\ln(O:E)_i)$	
Calibration-in-the-large (a)	$\ln\left(-\exp(a)P_O+\exp(a)+P_O\right)$	$\frac{(P_O-1)^2\left((P_O)^2+1\right)\left(\exp(P_O+a)\right)^2}{\left(P_O(-\exp(a))+P_O+\exp(a)\right)^2}Var(a)$	[Appendix 5]
Calibration plot	Extract N, P_O and P_E for each risk stratum.	See $Var(\ln(O:E))$ when *O* and *E* are known across different risk strata.	
Mean subject characteristics in the validation sample	Calculate P_E by incorporating the mean values of the subject characteristics in the prediction model. Afterwards, combine P_E and P_O to obtain O:E.	See $Var(\ln(O:E))$ when *O* and *E* are known.	

with $O:E_{lb}$ the lower and $O:E_{ub}$ the upper bound of the 95% confidence interval of the total O:E ratio and p the P-value of the total O:E ratio.

	Sergeant		Bridgewater		Calafiore		Karabulut		Toumpoulis		Biancari		Yap		Youn		Hirose	
LP	*N*	*O*	*N*†	*O*	*N*	*O*	*N*	*O*	*N*	*O*	*N*	*O*	*N*	*O*	*N*	*O*	*N*	*O*
0	191	0	1250	6	0	0	138	1.6	203.3	1.0	204	10.4	651.7	2.7	67	0.3	253	0
1	133	0	1400	10	0	0	138	1.6	203.3	1.0	164	9.5	651.7	2.7	67	0.3	236	0
2	225	2	1375	10	0	0	138	1.6	203.3	1.0	176	6.7	651.7	2.7	67	0.3	256	1
3	245	4	1350	14	0	0	121.7	1.7	493	5.4	129	4.6	665.3	5.7	103	1	258	0
4	209	1	1100	15	0	0	121.7	1.7	493	5.4	79	7.3	665.3	5.7	103	1	192	1
5	234	6	800	16	0	0	121.7	1.7	493	5.4	39	6.8	665.3	5.7	103	1	154	1
6	221	7	520	12	243	7.7	12.1	0.4	366	12.8	24	4.1	96.5	6.6	60.3	1.3	64	1
7	162	5	300	14	243	7.7	12.1	0.4	366	12.8	14	3.7	96.5	6.6	60.3	1.3	51	7
8	126	4	200	14	243	7.7	12.1	0.4	366	12.8	0.7	0.3	96.5	6.6	60.3	1.3	38	1
9	97	11	85	4	79.3	5.7	12.1	0.4	150.7	9.04	0.7	0.3	96.5	6.6	4.7	0.1	20	2
10	64	7	43	3	79.3	5.7	12.1	0.4	150.7	9.04	0.7	0.3	96.5	6.6	4.7	0.1	2.3	0.4
11	50	3	21	8	79.3	5.7	12.1	0.4	150.7	9.04	0.7	0.3	96.5	6.6	4.7	0.1	2.3	0.4
12	31	8	11	3	4.8	0.5	12.1	0.4	34.3	4.7	0.7	0.3	96.5	6.6	4.7	0.1	2.3	0.4
13	15	5	7	0	4.8	0.5	12.1	0.4	34.3	4.7	0.7	0.3	96.5	6.6	4.7	0.1	2.3	0.4
14	21	6	7	4	4.8	0.5	12.1	0.4	34.3	4.7	0.7	0.3	96.5	6.6	4.7	0.1	2.3	0.4
15	7	3	11	5	4.8	0.5	12.1	0.4	2.1	0.5	0.7	0.3	96.5	6.6	4.7	0.1	2.3	0.4
16	8	4	7	3	4.8	0.5	12.1	0.4	2.1	0.5	0.7	0.3	96.5	6.6	4.7	0.1	2.3	0.4
17	2	0	7	2	4.8	0.5	12.1	0.4	2.1	0.5	0.7	0.3	96.5	6.6	4.7	0.1	2.3	0.4
18	3	2	0	0	4.8	0.5	12.1	0.4	2.1	0.5	0.7	0.3	96.5	6.6	4.7	0.1	2.3	0.4
19	1	1	5	1	4.8	0.5	12.1	0.4	2.1	0.5	0.7	0.3	96.5	6.6	4.7	0.1	2.3	0.4
20	4	2	0	0	4.8	0.5	12.1	0.4	2.1	0.5	0.7	0.3	96.5	6.6	4.7	0.1	2.3	0.4
21	0	0	0	0	4.8	0.5	12.1	0.4	2.1	0.5	0.7	0.3	96.5	6.6	4.7	0.1	2.3	0.4
22	2	0	0	0	4.8	0.5	12.1	0.4	2.1	0.5	0.7	0.3	96.5	6.6	4.7	0.1	2.3	0.4

LP = EuroSCORE. Note that the scores for the risk factors in the EuroSCORE are added to give an approximate percentage predicted mortality, such that $E \approx LP \times N/100$

† Values were approximated from a histogram

9 Statistical models for meta-analysis

Below, we present the random effects meta-analysis models for summarizing estimates of model discrimination and calibration, and investigating sources of heterogeneity. For meta-analysis of the c-statistic we have:

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

with $logit\left(c_i\right)$ the logit of the c-statistic in the i[th] study, and $Var\left(logit\left(c_i\right)\right)$ its error variance which is assumed known. For meta-analysis of the total O:E ratio we have:

$$\ln\left(O:E_i\right) \sim N\left(\mu_{cal}, Var\left(\ln\left(O:E_i\right)\right) + \tau^2_{cal}\right)$$

with $\ln\left(O:E_i\right)$ the log of the total O:E ratio in the i[th] study, and $Var\left(\ln\left(O:E_i\right)\right)$ its error variance. The weighted average (e.g. of the logit c-statistic) is then given by μ and the extent of between-study heterogeneity is quantified by τ. We can back-transform the weighted averages into a summary c-statistic and total O:E ratio by applying $1/\left(1+\exp\left(\hat{\mu}_{discr}\right)\right)$ and, respectively, $\left(\hat{\mu}_{cal}\right)$.

When a meta-analysis is affected by heterogeneity, it is often helpful to calculate a prediction interval. This interval provides a range for the predicted model performance in a new validation of the model.[13] A 95% prediction interval for the c-statistic in a new setting is approximately given as

$$\frac{1}{1+\exp\left(-\hat{\mu}_{discr} \pm t_{n-2}\sqrt{\hat{\tau}^2_{discr} + SE\left(\hat{\mu}_{discr}\right)^2}\right)}$$

and a 95% prediction interval for the total O:E ratio in a new setting as

$$\exp\left(\hat{\mu}_{cal} \pm t_{n-2}\sqrt{\hat{\tau}^2_{cal} + SE\left(\hat{\mu}_{cal}\right)^2}\right)$$

In these equations, t_{n-2} is the $100\left(1-\alpha/2\right)$ percentile of the t distribution with $n-2$ degrees of freedom, where α is usually chosen as 0.05, to give a 5% significance level and thus 95% prediction interval.

We can extend aforementioned meta-analysis models to investigate whether the weighted average is influenced by study-level or summarized patient-level characteristics (e.g. mean age). The resulting models are also known as meta-regression models. For meta-regression of the c-statistic we have:

$$\begin{aligned} logit(c_i) &\sim N\left(\mu_{discr}, Var\left(logit(c_i)\right)+\tau^2_{discr}\right) \\ \mu_{discr} &= \alpha_{discr}+\beta_{discr}\,x_i \end{aligned}$$

where x_i indicates the explanatory or independent variable of the i[th] study. In the empirical example, for instance, we may use the standard deviation of the additive EuroSCORE in each validation study as values for x_i. The estimate for β_{discr} (and its standard error) then describes whether the weighted average of the logit c-statistic is modified by the explanatory variable x_i.

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

$$\begin{aligned} \ln(O:E_i) &\sim N\left(\mu_{cal}, Var\left(\ln(O:E_i)\right)+\tau^2_{cal}\right) \\ \mu_{cal} &= \alpha_{cal}+\beta_{cal}\,x_i \end{aligned}$$

Note that the interpretation of μ_{discr} and μ_{cal} is now dependent on the magnitude of x_i. It is therefore often helpful to transform the explanatory variable such that its mean equals zero.

Finally, it is also possible to jointly evaluate these estimates by performing a multivariate meta- analysis.[8] This may help to increase precision of summary estimates, and to avoid exclusion of studies for which relevant estimates are missing (e.g. because they were not reported). Multivariate meta-analysis may also help to obtain joint ranges of predictive performance, and to quantify the overall probability of 'good' performance in new populations. When jointly summarizing the c- statistic and the total O:E ratio, the meta-analysis model can be written as follows:

$$\begin{pmatrix} logit(c_i) \\ \ln(O:E_i) \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_{discr} \\ \mu_{cal} \end{pmatrix}, S_i+T\right)$$

with

$$S_i=\begin{pmatrix} Var\left(logit(c_i)\right) & Cov\left(logit(c_i), \ln(O:E_i)\right) \\ Cov\left(logit(c_i), \ln(O:E_i)\right) & Var\left(\ln(O:E_i)\right) \end{pmatrix}$$

and

$$T=\begin{pmatrix} \tau^2_{discr} & \rho\,\tau_{discr}\,\tau_{cal} \\ \rho\,\tau_{discr}\,\tau_{cal} & \tau^2_{cal} \end{pmatrix}$$

where MVN denotes a multivariate normal distribution, S_i the within-study variance-covariance matrix of the i[th] study and T the between-study variance-covariance matrix. In contrast to the standard (univariate) meta-analysis model, this model takes the within-study and between-study correlation between the c-statistic and total O:E ratio into account when deriving summary estimates for discrimination and calibration performance. As a result, it becomes possible to 'borrow' information across studies when some entries of $\text{logit}(c_i)$ or $\ln(\text{O: E}_i)$ are unknown. Because estimates for within-study covariance are often not reported, one may assume that $Cov\left(logit\left(c_i\right), \ln\left(O:E_i\right)\right)=0$ A motivation for this is given below.

For logistic regression models, the total O:E ratio and calibration-in-the-large can be written as a function of the outcome prevalence (P_O) and the average linear predictor (LP) in the validation sample. The c-statistic is not related to any of these statistics, but depends on the deviation of linear predictor between patients that experience and do not experience the outcome of interest.[3,16] As a result, for validation of a logistic regression model, the c-statistic should be independent from the total O:E ratio (or from the calibration-in-the-large). This phenomenon has also been suggested by Prof. Steyerberg. An example can be found in,[14] where the within-study correlation between ln(E:O) and the c-statistic was calculated in 12 studies using the corresponding individual participant data. Results in their supplementary material 3(b) indicate that these correlations were very close to 0.

10 Investigating heterogeneity in the performance of the additive EuroSCORE

To investigate whether population differences generated heterogeneity across the validation studies, we performed meta-regression analyses and implemented the Hartung-Knapp-Sidik-Jonkman method for deriving confidence intervals. We examined the following explanatory variables from the external validation studies in separate meta-regression models: the spread of the additive EuroSCORE, the spread of participant age, the mean EuroSCORE value, the calendar year of study recruitment and the continent in which the validation study was conducted. For all models, we standardized the (continuous) explanatory variable by subtracting their mean value.

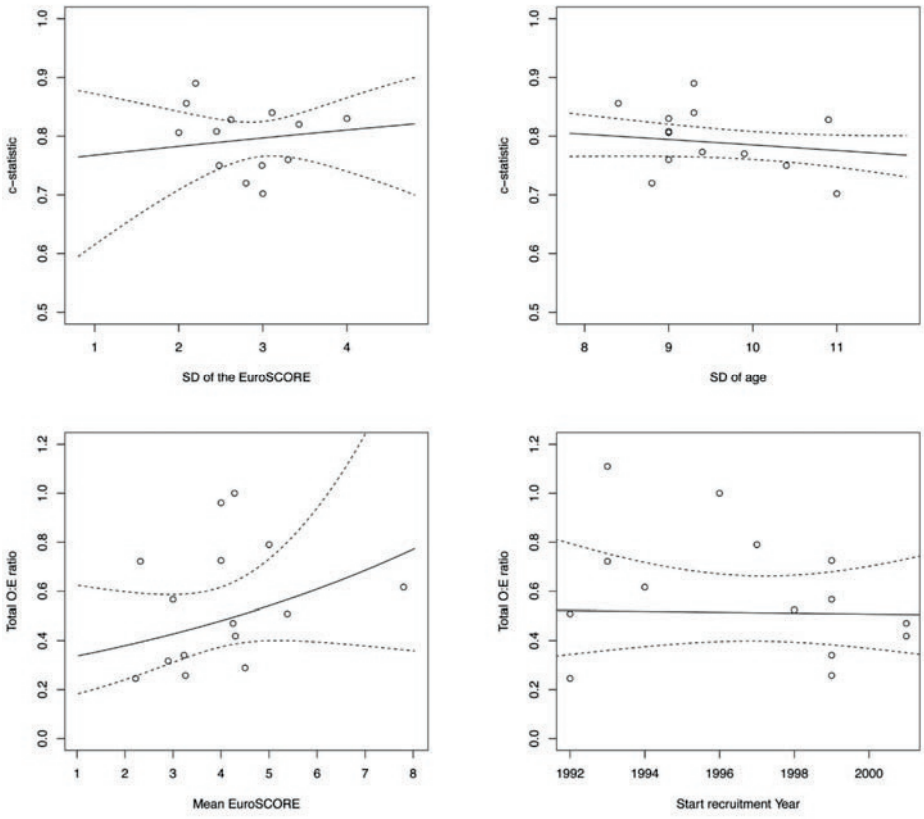


Figure 2: Results from random-effects meta-regression models. Dashed lines indicate the bounds of the 95% confidence interval around the regression line. Dots indicate the included validation studies.

Table A.3 describes the total O:E ratio across different continents, and can be used to assess absolute differences in predicted risks. For instance, the median mortality rate of European validation studies is 1.7%. In this population, EuroSCORE yields a total O:E ratio of 0.55. Hence, we have: $0.017/P_E=0.55$ such that $P_O-P_E=0.017-0.031=-0.014=-1.4\%$. Evidently, it is also possible to conduct a proper meta-analysis on the absolute difference O minus E. In that case, no transformations are needed for conducting the meta-analysis.

Continent		Mortality		Total O:E ratio	
	N	Median	IQR	Summary	95% CI
Overall	18	1.8%	1.3% – 3.2%	0.53	0.42 – 0.67
Europe	10	1.7%	1.2% – 2.9%	0.55	0.40 – 0.75
South-East Asia	4	1.7%	1.2% – 2.2%	0.42	0.25 – 0.69
North America	2	2.1%	2.0% – 2.2%	0.46	0.23 – 0.93
South America & South Africa	2	4.3%	4.0% – 4.6%	0.95	0.42 – 2.14

Table A.3: Results from random-effects meta-regression analysis where the total O:E ratio is adjusted for continent.

IQR = Interquartile range; CI = confidence interval; Europe = subjects from the United Kingdom, Italy, Turkey, Belgium, Sweden and/or Finland; South-East Asia = subjects from China, Japan, Australia or Korea; South America = subjects from Brazil.

To further investigate the extent of mis-calibration, we used the reported calibration tables and histograms within the primary validation studies to investigate the total O:E ratio of the additive EuroSCORE across different subgroups. For each risk stratum, the log of the proportion of observed events were pooled using random effects meta-analysis. Results in Figure 3 again demonstrate that the EuroSCORE tends to over-estimate the risk of early mortality in low-risk subgroups.

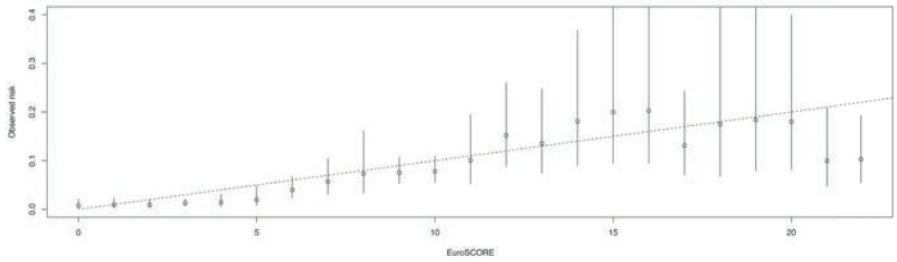


Figure 3: Results from the subgroup analyses in the external validation studies (*n* = 9). Estimates below the dashed reference line indicate that the additive EuroSCORE over-estimated the occurrence of early mortality in the corresponding subgroup.

References

1. Altman DG, Bland JM. How to obtain the confidence interval from a P value. *BMJ* 2011;343:d2090.
2. Altman DG, Bland JM. How to obtain the P value from a confidence interval. *BMJ* 2011;343:d2304.
3. Austin PC, Pencinca MJ, Steyerberg EW. Predictive accuracy of novel risk factors and markers: A simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model. *Stat Methods Med Res* 2017;26(3):1053-77.
4. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol* 2012;12:82.
5. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;353:i2416.
6. Dorfman R. A note on the delta-method for finding variance formulae. *The Biometric Bulletin* 1938;1(129-137):92.
7. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29-36.
8. Jackson D, Riley RD. A refined method for multivariate meta-analysis and meta-regression. *Stat Med* 2014;33(4):541-54.
9. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11(10):e1001744.
10. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg* 1999;16(1):9-13.
11. Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: asymptotic methods and evaluation. *Stat Med* 2006;25(4):559-73.
12. Newson R. Parameters behind" nonparametric" statistics: Kendall's tau, Somers' D and median differences. 2002.
13. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011;342:d549.
14. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2015.

15. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* 2009;302(21):2345-52.
16. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010;172(8):971-80.

Chapter 3

Prediction models for cardiovascular disease risk in the general population: systematic review

Johanna AAG Damen
Lotty Hooft
Ewoud Schuit
Thomas PA Debray
Gary S Collins
Ioanna Tzoulaki
Camille M Lassale
George CM Siontis
Virginia Chiocchia
Corran Roberts
Michael Maia Schlüssel
Stephen Gerry
James A Black
Pauline Heus
Yvonne T van der Schouw
Linda M Peelen
Karel GM Moons

BMJ. 2016;353:i2416

Abstract

Objective: To provide an overview of prediction models for risk of cardiovascular disease (CVD) in the general population.

Design: Systematic review.

Data sources: Medline and Embase until June 2013.

Eligibility criteria for study selection: Studies describing the development or external validation of a multivariable model for predicting CVD risk in the general population.

Results: 9965 references were screened, of which 212 articles were included in the review, describing the development of 363 prediction models and 473 external validations. Most models were developed in Europe (n=167, 46%), predicted risk of fatal or non-fatal coronary heart disease (n=118, 33%) over a 10 year period (n=209, 58%). The most common predictors were smoking (n=325, 90%) and age (n=321, 88%), and most models were sex specific (n=250, 69%). Substantial heterogeneity in predictor and outcome definitions was observed between models, and important clinical and methodological information were often missing. The prediction horizon was not specified for 49 models (13%), and for 92 (25%) crucial information was missing to enable the model to be used for individual risk prediction. Only 132 developed models (36%) were externally validated and only 70 (19%) by independent investigators. Model performance was heterogeneous and measures such as discrimination and calibration were reported for only 65% and 58% of the external validations, respectively.

Conclusions: There is an excess of models predicting incident CVD in the general population. The usefulness of most of the models remains unclear owing to methodological shortcomings, incomplete presentation, and lack of external validation and model impact studies. Rather than developing yet another similar CVD risk prediction model, in this era of large datasets, future research should focus on externally validating and comparing head-to-head promising CVD risk models that already exist, on tailoring or even combining these models to local settings, and investigating whether these models can be extended by addition of new predictors.

Introduction

Cardiovascular disease (CVD) is a leading cause of morbidity and mortality worldwide,[3] accounting for approximately one third of all deaths.[4] Prevention of CVD requires timely identification of people at increased risk to target effective dietary, lifestyle, or drug interventions. Over the past two decades, numerous prediction models have been developed, which mathematically combine multiple predictors to estimate the risk of developing CVD—for example, the Framingham,[5-7] SCORE,[8] and QRISK[9-11] models. Some of these prediction models are included in clinical guidelines for therapeutic management[12] and are increasingly advocated by health policymakers. In the United Kingdom, electronic health patient record systems now have QRISK2 embedded to calculate 10 year CVD risk.

Several reviews have shown that there is an abundance of prediction models for a wide range of CVD outcomes.[16] However, the most comprehensive review[16] includes models published more than 10 years ago (search carried out in 2003). More recent reviews have shown that the number of published prediction models has increased dramatically since then; furthermore, these reviews have not systematically described the outcomes that the models intended to predict, the most common predictors, the predictive performance of all these models, and which developed prediction models have been externally validated.[18,19]

We carried out a systematic review of multivariable prediction models developed to predict the risk of developing CVD in the general population, to describe the characteristics of the models' development, included predictors, CVD outcomes predicted, presentation, and whether they have undergone external validation.

Methods

We conducted our systematic review following the recently published guidance from the Cochrane Prognosis Methods Group, using the CHARMS checklist, for reviews of prediction model studies.[13]

Literature search

We performed a literature search in Medline and Embase on 1 June 2013 using search terms to identify primary articles reporting on the development and/or validation of models predicting incident CVD, published from 2004 onwards (see supplementary table 1). Articles published before 2004 were identified from a previously published comprehensive systematic review,[16] and a cross reference check was performed for all reviews on CVD prediction models identified by our search. For external validation studies

where the development study was not identified by our search, we manually retrieved and included in the review the original article describing the development of the model.

Eligibility criteria

We included all primary articles that reported on one or more multivariable (that is, including at least two predictors[20]) prediction models, tools, or scores, that have been proposed for individual risk estimation of any future CVD outcome in the general population. We differentiated between articles reporting on the development[21-23] or external validation[23-25] of one or more prediction models (box 1).[1,2,14,17] Studies reporting on the incremental value or model extension—that is, evaluating the incremental value of one or more new predictors to existing models,[26] were excluded. We classified articles as development studies if they reported the development of a model in their objectives or conclusions, or if it was clear from other information in the article that they developed a prediction model for individual risk estimation (eg, if they presented a simplified risk chart). Included articles had to report original research (eg, reviews and letters were excluded), study humans, and be written in English. Articles were included if they reported models for predicting any fatal or non-fatal arterial CVD event. We excluded articles describing models for predicting the risk of venous disease; validation articles with a cross sectional study design that, for example, compared predicted risks of two different models at one time point without any association with actual CVD outcomes; and articles describing models developed from or validated exclusively in specific diseased (patient) populations, such as patients with diabetes, with HIV, with atrial fibrillation, or undergoing any surgery. Furthermore, we excluded methodological articles and articles for which no full text was available through a license at our institutes. Impact studies identified by our search were excluded from this review but were described in a different review.[27] External validation articles were excluded if the corresponding development article was not available.

A single article can describe the development and/or validation of several prediction models, and the distinction between models is not always clear. We defined reported models as separate models whenever a combination of two or more predictors with unique predictor-outcome association estimates were presented. For example, if a model was fitted after stratification for men and women yielding different predictor-outcome associations (that is, predictor weights), we scored it as two separate models. Additionally, two presented models yielding the same predictor-outcome associations but with a different baseline hazard or risk estimate, were considered separately.

Screening process

Initially pairs of two reviewers (JAB, TPAD, CML, LMP, ES, GCMS) independently screened retrieved articles for eligibility on title and subsequently on abstract. Disagreements were

resolved by iterative screening rounds. After consensus, full text articles were retrieved and one reviewer (JAB, GSC, VC, JAAGD, SG, TPAD, PH, LH, CML, CR, ES, GCMS, MMS, IT) screened the full text articles and extracted data. In case of doubt, a second (JAAGD or GSC) or third (ES or KGMM) reviewer was involved.

Box 1: Definitions of technical terms

Internal validation—testing a model's predictive accuracy by reusing (parts of) the dataset on which the model was developed. The aim of internal validation is to assess the overfit and correct for the resulting "optimism" in the performance of the model. Examples are cross validation and bootstrapping[1]

External validation—testing a model's predictive accuracy in a population other than the development population[2]

Prediction horizon—time frame for which the model is intended to predict the outcome[13]

Discrimination—ability of the model to distinguish between people who do and do not develop the outcome of interest[14]

Calibration—agreement between predicted and observed numbers of events[1,15]

Updating—adjusting a previously developed model to a new setting or study population, to improve model fit in that population. Several forms of updating exist, including intercept recalibration, slope recalibration, and refitting all coefficients of a model.[17-19] It is also possible to combine and update existing models

Data extraction and critical appraisal

We categorised the eligible articles into two groups: development articles, and external validation (with or without model recalibration) articles.

The list of extracted items was based on the recently issued Cochrane guidance for data extraction and critical appraisal for systematic reviews of prediction models (the CHARMS checklist[13]) supplemented by items obtained from methodological guidance papers and previous systematic reviews in the specialty.[13,28-31] The full list of extracted items is available on request. Items extracted from articles describing model development included study design (eg, cohort, case-control), study population, geographical location, outcome, prediction horizon, modelling method (eg, Cox proportional hazards model, logistic model), method of internal validation (eg, bootstrapping, cross validation), number of study participants and CVD events, number and type of predictors, model presentation (eg, full regression equation, risk chart), and predictive performance measures (eg, calibration, discrimination). For articles describing external validation of a prediction model we extracted the type of external validation (eg, temporal, geographical[25,32]), whether or not the validation was performed by the same investigators who developed the model, study population, geographical location, number of participants and events, and the model's performance before and (if conducted) after model recalibration. If an article described multiple models, we carried out separate data extraction for each model.

To accomplish consistent data extraction, a standardised data extraction form was piloted and modified several times. All reviewers were extensively trained on how to use the form. A second reviewer (JAAGD) checked extracted items classed as "not reported" or "unclear," or unexpected findings. We did not explicitly perform a formal risk of bias assessment as no such tool is currently available for studies of prediction models.

Descriptive analyses

Results were summarised using descriptive statistics. We did not perform a quantitative synthesis of the models, as this was beyond the scope of our review, and formal methods for meta-analysis of prediction models are not yet fully developed.

Patient involvement

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

Results

The search strategy identified 9965 unique articles, of which 8577 were excluded based on title and abstract. In total, 1388 full texts were screened, of which 212 articles met the eligibility criteria and were included in this review (Figure 1). In total, 125 articles concerned the development of one or more CVD risk prediction models and 136 articles described the external validation of one or more of these models (see supplementary table 2). Frequently, articles described combinations of development or external validation (Figure 1), therefore the total number does not sum up to 212. The number of development and external validation studies increased over time (Figure 2).

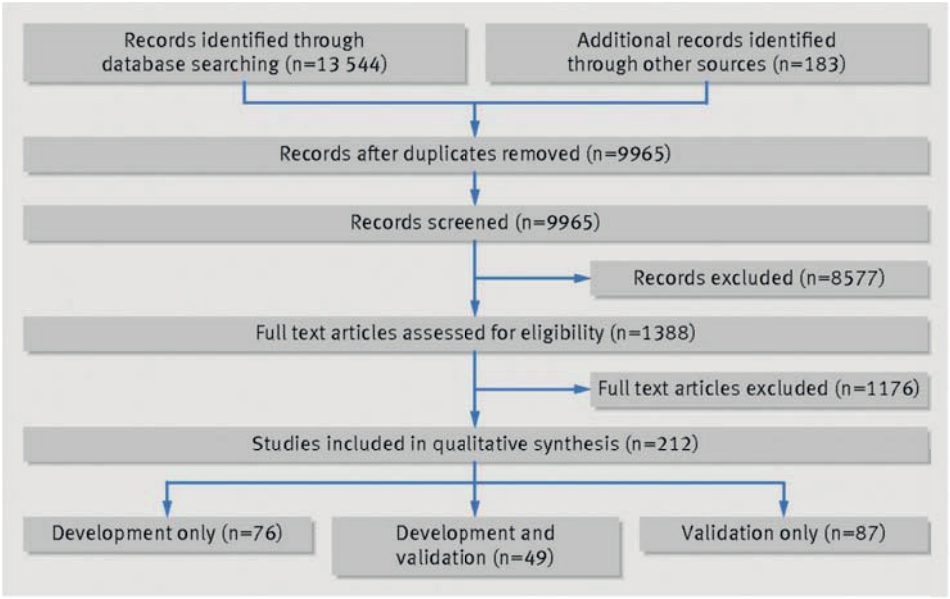


Figure 1: Flow diagram of selected articles

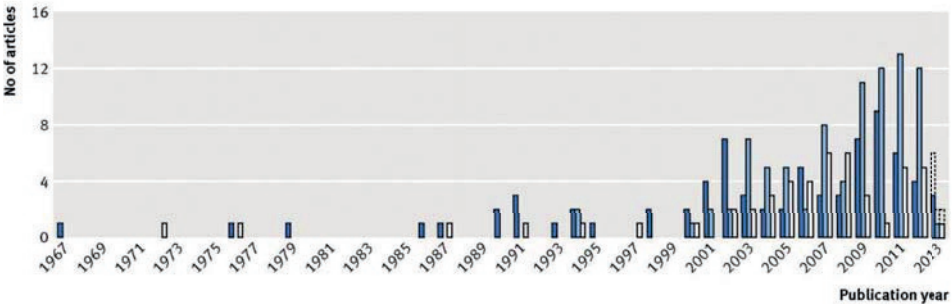


Figure 2: Numbers of articles in which only one or more models were developed (dark blue), only one or more models were externally validated (light blue), or one or more models were developed and externally validated (white), ordered by publication year (up to June 2013). Predictions of the total numbers in 2013 are displayed with dotted lines

Studies describing the development of CVD prediction models

Study designs and study populations

Overall, 125 articles described the development of 363 different models. Most of the prediction models (n=250, 69%) were developed using data from a longitudinal cohort study (see supplementary figure 1A); most originated from Europe (n=168, 46%) or the United States and Canada (n=132, 36%, see supplementary figure 1B). No models were developed using data from Africa. Several cohorts were used multiple times for model development—for example, the Framingham cohort, yielding 69 models in 23 papers.

Study populations (that is, case mix) differed noticeably between studies, mainly for age, sex, and other patient characteristics. Most models were developed for people with ages ranging from 30 to 74 years (n=206, 57%), although 69 different age ranges were reported (see supplementary figure 1C). The majority of models was sex specific (men n=142, 39%; women n=108, 30%), and for most models (n=230, 63%), investigators explicitly stated they excluded study participants with existing CVD (including coronary heart disease, stroke, other heart diseases, or combinations of those), or with other diseases such as cancer (n=21, 6%) or diabetes (n=43, 12%).

CVD outcomes

We observed large variation in predicted outcomes. Although the majority of prediction models focused on (fatal or non-fatal) coronary heart disease or CVD (n=118, 33% and n=95, 26%), 19 other outcomes were identified, such as (fatal or non-fatal) stroke, myocardial infarction, and atrial fibrillation (see supplementary table 3). On top of this, the definitions of these outcomes showed considerable heterogeneity, with, for example, more than 40 different definitions for fatal or non-fatal coronary heart disease (see supplementary table 4). International classification of disease codes were specified for 82 out of 363 models (23%).

Predictors

The median number of predictors included in the developed models was 7 (range 2-80). In total, more than 100 different predictors were included (Figure 3). Sex was included in 88 (24%) models; however, 250 (69%) models were explicitly developed only for men or only for women. Most of the models (n=239, 66%) included a set of similar predictors, consisting of age, smoking, blood pressure, and blood cholesterol measurements. Other prevalently selected predictors were diabetes (n=187, 52%) and body mass index (n=107, 29%). Treatment modalities were included in a few prediction models; 56 models (15%) included use of antihypertensive treatment and no models included use of lipid lowering drugs.

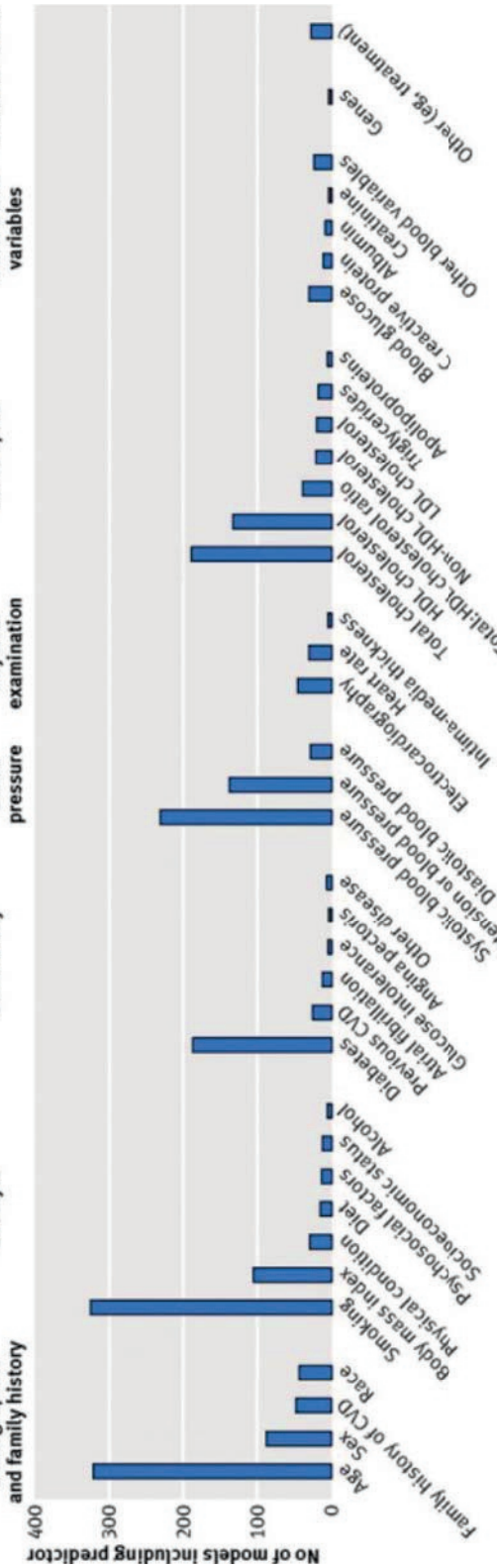


Figure 3: Main categories of predictors included in developed models. CVD=cardiovascular disease; HDL=high density lipoprotein; LDL=low density lipoprotein

Sample size

The number of participants used to develop the prediction models ranged from 51 to 1 189 845 (median 3969), and the number of events ranged between 28 and 55 667 (median 241). The number of participants and the number of events were not reported for 24 (7%) and 74 (20%) models, respectively. The number of events for each variable included in the final prediction model could be calculated for 252 (69%) models and ranged from 1 to 4205. For 25 out of these 252 (10%) models, this number of events for each variable was less than 10.[33,34]

Modelling method and prediction horizon

We found that most prediction models were developed using Cox proportional hazards regression (n=160, 44%), accelerated failure time analysis (n=77, 21%), or logistic regression (n=71, 20%). For 36 models (10%) the method used for statistical modelling was not clear (see supplementary table 5). The prediction horizon ranged between 2 and 45 years, with the majority of studies predicting CVD outcomes for a five year or 10 year horizon (n=47, 13% and n=209, 58%, respectively). For 49 models (13%), the prediction horizon was not specified (see supplementary table 6).

Model presentation

For 167 models (46%) the complete regression formula, including all regression coefficients and intercept or baseline hazard, were reported. Of the other 196 models, 104 (53%) were presented as online calculator, risk chart, sum score, or nomogram to allow individual risk estimation. For the remaining models (n=92, 25%) insufficient information was presented to allow calculation of individual risks.

Predictive performance

At least one measure of predictive performance was reported for 191 of the 363 (53%) models (Table 1). For 143 (39%) models, discrimination was reported as a C statistic or area under the receiver operating characteristic curve (range 0.61 to 1.00). Calibration was reported for 116 (32%) models, for which a variety of methods was used, such as a Hosmer-Lemeshow test (n=60, 17%), calibration plot (n=31, 9%) or observed:expected ratio (n=12, 3%). For 99 (27%) models, both discrimination and calibration were reported. Table 2 shows that reporting of discriminative performance measures seems to have increased over time, whereas reporting of calibration seems to remain limited.

Table 1: Performance measures reported for developed models. Values are numbers (percentages) unless stated otherwise

Performance measures	Development	Validation
Discrimination measures:		
C statistic/AUC	143 (39)	303 (64)
D statistic	5 (1)	45 (9)
Other*	24 (7)	8 (2)
Any	163 (45)	306 (65)
Calibration measures:		
Plot	31 (9)	122 (26)
Table	34 (9)	62 (13)
Slope	3 (1)	7 (1)
Intercept	2 (1)	7 (1)
Hosmer Lemeshow test	60 (17)	68 (14)
Observed:expected ratio	12 (3)	124 (26)
Other†	7 (2)	20 (4)
Any	116 (32)	277 (58)
Overall performance measures:		
R^2	13 (4)	49 (10)
Brier score	15 (4)	45 (9)
Other‡	10 (3)	1 (<0.5)
Any	35 (10)	68 (14)
Any performance measure	191 (53)	398 (84)
Total	363	474

AUC=area under receiver operating characteristic curve. Numbers add up to over 363 since papers may have reported more than one predictive performance measure. *For example, sensitivity, specificity. †For example, Grønnesby-Borgan χ^2 test. ‡For example, Akaike information criterion, bayesian information criterion.

Table 2: Reporting of performance measures for models across years of publication. Values are numbers (percentages) unless stated otherwise

Performance measures	Publication year			
	1967-2001	2002-05	2006-08	2009-13
Development:				
Discrimination	12 (14)	46 (55)	41 (44)	64 (64)
Calibration	13 (15)	41 (49)	25 (27)	37 (37)
Overall performance*	0 (0)	2 (2)	12 (13)	21 (21)
Any performance	25 (29)	48 (58)	42 (45)	76 (76)
Total	87	83	93	100
Validation:				
Discrimination	12 (32)	41 (44)	71 (68)	182 (77)
Calibration	29 (76)	45 (48)	64 (61)	139 (59)
Overall performance	0 (0)	0 (0)	22 (21)	46 (19)
Any performance	31 (82)	56 (60)	98 (93)	213 (90)
Total	38	93	105	237

*Performance measures giving overall indication of goodness of fit of a model, such as R^2 and brier score.[35]

Internal validation

In total, 80 of the 363 developed models (22%) were internally validated, most often using a random split of the dataset (n=27), bootstrapping (n=23), or cross validation (n=22).

Studies describing external validation of a prediction model

In 136 articles, 473 external validations were performed. However, the majority of the 363 developed models (n=231, 64%) has never been externally validated. Out of the 132 (36%) models that were externally validated, 35 (27%) were validated once, and 38 (29%) (originally developed and described in seven articles) were validated more than 10 times. The most commonly validated models were Framingham (Wilson 1998, n=89),[7] Framingham (Anderson 1991, n=73),[5] SCORE (Conroy 2003, n=63),[8] Framingham (D'Agostino 2008, n=44),[36] Framingham (ATP III 2002, n=31),[37] Framingham (Anderson 1991, n=30),[6] and QRISK (Hippisley-Cox 2007, n=12)[10] (Table 3).

Table 3: List of the models that were validated at least three times, and their predicted outcomes (sorted by number of validations)

Reference (No of developed models)	Predicted outcomes	No of validations
Framingham Wilson 1998[7] (n=2*)	Fatal or non-fatal CHD	89
Framingham Anderson 1991[5] (n=12)	Fatal or non-fatal: CHD, CVD, myocardial infarction, and stroke	73
SCORE Conroy 2003[8] (n=12)	Fatal: CHD, CVD, and non-CHD	63
Framingham D'Agostino 2008[36] (n=4)	Fatal CVD	44
Framingham ATP III 2002[37] (n=2)	Fatal or non-fatal CHD	31
Framingham Anderson 1991[6] (n=4)	Fatal or non-fatal CHD	30
QRISK Hippisley-Cox 2007[10] (n=2)	Fatal CVD	12
PROCAM Assman 2002[38] (n=1)	Fatal or non-fatal CHD	8
Framingham Wolf 1991[39] (n=2)	Fatal or non-fatal stroke	8
Chambless 2003[40] (n=4)	Fatal or non-fatal CHD	7
Friedland 2009[41] (n=7)	Fatal or non-fatal: CHD, myocardial infarction, and stroke; claudication; coronary artery bypass grafting; percutaneous transluminal coronary angioplasty; transient ischaemic attack	6
QRISK Hippisley-Cox 2010[9] (n=2)	Fatal CVD	6
Keys 1972[42] (n=4)	Fatal or non-fatal CHD	6
Leaverton 1987[43] (n=4)	Fatal CHD	6
Asia Pacific cohort studies 2007[44] (n=4)	Fatal CVD	4
Woodward 2007[45] (n=2)	Fatal CVD	4
Levy 1990[46] (n=4)	Fatal or non-fatal CHD	4
Chien 2012[47] (n=3)	Fatal or non-fatal CHD	3
Framingham unspecified†	—	32

CHD=coronary heart disease; CVD=cardiovascular disease. *Number of models developed in this article. †Authors stated they externally validated the Framingham model without referencing the specific model.

Out of the 132 externally validated models, 45 (34%) were solely externally validated in the same paper in which their development was described, 17 (13%) were externally validated in a different paper but with authors overlapping between the development and validation paper, and 70 (53%) were validated by independent researchers. Sample sizes of the validation studies ranged from very small (eg, 90 participants or one event) to very large (eg, 1 066 127 participants or 51 340 events). Most external validations were performed in a different geographical area from the development study—for example, the Framingham (Anderson 1991)[5] model (developed on data from the United States) was often validated outside North America, namely in Europe (71% of its validations), Australia (16%), or Asia (4%) (Table 4). There was considerable heterogeneity in eligibility criteria for patients between validation and development studies. For example, for the seven aforementioned models, 13% of the validation studies were performed in the same age range for which the model was originally developed. For Framingham (Anderson 1991)[5] only few (n=12, 16%) validations were performed in people outside these age ranges, whereas for Framingham (Wilson 1998)[7] and SCORE (Conroy 2003)[8] this happened more often (n=34, 38% and n=33, 52%, respectively; see supplementary figure 2).

In external validation studies, the C statistic was reported for 303 (64%) models. For 277 models (58%) a calibration measure was reported by using a calibration plot (n=122, 26%), an observed:expected ratio (n=124, 26%), the Hosmer-Lemeshow test (n=68, 14%), a calibration table (that is, a table with predicted and observed events; n=62, 13%), or a combination of those (Table 1). Both discrimination and calibration were reported for 185 (39%) external validations. The discriminative ability and calibration of the three most often validated models (Framingham (Wilson 1998),[7] Framingham (Anderson 1991),[5] and SCORE (Conroy 2003[8])) varied between validation studies, with C statistics between 0.57 and 0.92, 0.53 and 0.99, and 0.62 and 0.91, respectively, and observed:expected ratios between 0.37 and 1.92, 0.18 and 2.60, and 0.28 and 1.50, respectively (table 4).

Models that were external validated differed in many respects from the non-validated models (see supplementary table 7). Ninety three per cent of validated models were developed using longitudinal cohort data versus 81% of non-validated models, 34% versus 15% were internally validated, and 83% versus 70% were presented in a way that allowed the calculation of individual risk. The median publication year for validated models was 2002 (or 2003 after excluding the earliest Framingham models) versus 2006 for models that were not validated. In addition, validated models were developed in studies with a median of 364 events versus 181 for non-validated models. More than half (75 out of 132, 57%) of the models developed in the United States or Canada were validated, compared with 24% (40 out of 168) of models developed from Europe and 16% (7 out of 43) from Asia; excluding the Framingham prediction models did not influence these percentages. None of the models developed in Asia was validated by independent researchers, whereas 41 out of 132 (31%) models from the United States and 26 out of 168 (15%) from Europe were validated by independent researchers.

Table 4: Description of study populations and design characteristics used to validate seven most often (>10 times, see table 3) validated models. Values are numbers (percentages) unless stated otherwise

Characteristics	Framingham Wilson 1998[7](n=89)†	Anderson 1991[5] (n=73)	SCORE: Conroy 2003[8] (n=63)	Framingham D'Agostino 2008[36] (n=44)	ATP III 2002[37](n=31)	Anderson 1991[6] (n=30)	QRISK: Hippisley-Cox 2007[10] (n=12)
Location:							
Asia	9 (10)	3 (4)	2 (3)	8 (18)	2 (6)	2 (7)	0 (0)
Australia	0 (0)	12 (16)	4 (6)	2 (5)	1 (3)	2 (7)	0 (0)
Europe	34 (38)	52 (71)	47 (75)	20 (45)	6 (19)	18 (60)	12 (100)
North America	46 (52)	6 (8)	10 (16)	14 (32)	22 (71)	8 (27)	0 (0)
Age:							
Same age range as development study*	2 (3)	21 (29)	4 (6)	5 (11)	0 (0)	0 (0)	12 (100)
Young people (<50 years)	3 (3)	6 (8)	4 (6)	3 (7)	3 (10)	1 (3)	0 (0)
Older people (>60 years)	5 (6)	7 (10)	4 (6)	3 (7)	10 (32)	0 (0)	0 (0)
Other	79 (89)	39 (53)	51 (81)	33 (25)	18 (58)	29 (97)	0 (0)
Sex:							
Men	38 (43)	30 (41)	23 (37)	11 (25)	10 (32)	16 (53)	6 (50)
Women	29 (33)	25 (34)	23 (37)	11 (25)	10 (32)	13 (43)	6 (50)
Men and women	22 (25)	18 (25)	17 (27)	22 (50)	11 (35)	1 (3)	0 (0)

Table 4: Continued

Characteristics	Framingham Wilson 1998[7] (n=89)†	Anderson 1991[5] (n=73)	SCORE: Conroy 2003[8] (n=63)	Framingham D'Agostino 2008[36] (n=44)	ATP III 2002[37] (n=31)	Anderson 1991[6] (n=30)	QRISK: Hippisley-Cox 2007[10] (n=12)
Median (range) No of participants	2716 (100-163 627), n=87	2423 (262-797 373), n=71	8025 (262-44 649), n=63	2661 (272-542 987), n=44	3029 (534-36 517), n=31	3573 (331-542 783), n=30	536,400 (301,622-797 373), n=12
Median (range) No of events	146 (8-24 659), n=65	128 (1-42 408), n=59	224 (16-1722), n=54	164 (15-26 202), n=35	415 (35-2343), n=29	188 (4-26 202), n=28	29 057 (18 027-42 408), n=6
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	0.77 (0.58-0.84), n=28	0.66 (0.60-0.84), n=21	0.75 (0.63-0.78), n=6	0.79 (0.76-0.81), n=12
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	0.80 (0.62-0.96), n=3	0.47 (0.47-0.47), n=1	0.71 (0.32-3.92), n=14	0.94 (0.87-1.00), n=4

*30-74 (Framingham Wilson 1998,[7] Anderson 1991,[5,6] D'Agostino 2008,[36] ATP III 2002[37]), 40-65 (SCORE Conroy 2003[8]), 35-74 (QRISK Hippisley-Cox 2007[10]).
†Number of times model was externally validated. ‡Number of models for which this information was reported.

Discussion

This review shows that there is an abundance of cardiovascular risk prediction models for the general population. Previous reviews also indicated this but were conducted more than a decade ago,[16] excluded models that were not internally or externally validated,[18] or excluded articles that solely described external validation.[19]

Clearly, the array of studies describing the development of new risk prediction models for cardiovascular disease (CVD) in the general population is overwhelming, whereas there is a paucity of external validation studies for most of these developed models. Notwithstanding a few notable exceptions, including the Framingham and SCORE models, most of the models (n=231, 64%) have not been externally validated, only 70 (19%) have been validated by independent investigators, and only 38 (10%)—from only seven articles—were validated more than 10 times.

Healthcare professionals and policymakers are already in great doubt about which CVD prediction model to use or advocate in their specific setting or population. Instead of spending large amounts of research funding on the development of new models, in this era of large datasets, studies need to be aimed at validating the existing models and preferably using head-to-head comparisons of their relative predictive performance, be aimed at tailoring these models to local settings or populations, and focus on improving the predictive performance of existing models by the addition of new predictors.[48]

We found much variability in geographical location of both model development and model validation, but the majority of models were developed and validated in European and Northern American populations. Although the World Health Organization states that more than three quarters of all CVD deaths occur in low income and middle income countries,[49] a prediction model for people from Africa or South America has only recently been developed.[50] Several prediction models have been developed using data from Asia (eg,[44,51,52]) but none has yet been externally validated by independent researchers. Models tailored to these countries are important, as it is known that predictor-outcome associations vary among ethnic groups.[53]

With respect to outcome definitions, most models aimed to predict the risk of fatal or non-fatal coronary heart disease or the combined outcome of CVD. But we identified over 70 different definitions for these two outcomes. In addition, most outcomes were not fully defined and ICD codes were presented for only a few of the predicted outcomes. Without direct head-to-head comparison studies, these differences make it difficult to compare and choose between the existing prediction models based on our review, let alone to decide on which model to choose or advocate in a particular setting. Different definitions of CVD outcome lead to different estimated predictor effects, thus to different predicted probabilities and model performances, and consequently indicate different treatment strategies based on these prediction models. A more uniform definition and reporting of the predicted outcomes, preferably by explicit reporting of the ICD-9 or

ICD-10 codes for each outcome, would help the comparison of developed risk models, and their recommendation for and translation into clinical practice. Providing clear outcome definitions enhances not only the reporting of the development studies but also the conduct of external validation of developed models and, most importantly, the clinical implementation of the models by others.[30]

Most models (66%) were based on a common set of predictors, consisting of age, smoking, blood pressure, and cholesterol levels. Additional to this set, a large number (>100) of predictors have been included in models only once or twice. Interestingly, all these extended models have rarely been externally validated. This suggests that there is more emphasis placed on repeating the process of identifying predictors and developing new models rather than validating, tailoring, and improving existing CVD risk prediction models.

Strengths and limitations of this study

The major strengths of this review include the comprehensive search, careful selection of studies, and extensive data extraction on key characteristics of CVD risk prediction models, including the predictors, outcomes, and studied populations. However, this review also has some limitations. Firstly, we performed our search almost three years ago, and since then more than 4000 articles have been published that matched our search strategy. Therefore, some newly developed prediction models, such as the Pooled Cohort Equations[12] and GLOBORISK,[50] are not included in this overview. However, considering the large number of included models, including these articles is unlikely to change our main conclusions and recommendations. Moreover, it is this large number of newly identified articles in only two years, that actually underlines our main conclusions and reaffirms the necessity for changes regarding CVD risk prediction and a shift in focus from model development to model validation, head-to-head comparison, model improvement, and assessment of modelling impact. Secondly, we excluded articles not written in English (n=65) and for which no full text was available (n=124). This may have led to some underestimation of the number of models and external validations in the search period, and it might have affected the geographical representation. Thirdly, for external validations of a model published in an article in which several models were developed, it was often not stated exactly which of these models was validated. We therefore assumed all developed models in such articles as validated, which could even have resulted in an overestimation of the number of validated models.

Comparison with other studies

As with previous reviews in other specialties,[29,54,55] we found that important clinical and methodological information needed for validation and use of a developed model by others, was often missing. Incomplete reporting is highlighted as an important source of research waste, especially because it prevents future studies from summarising or

properly building on previous work, and guiding clinical management.[56] We have already dealt with the poor reporting of predicted outcome definitions and measurement. Although we observed an improvement in the reporting of discriminative performance measures over time, for 10% of the developed models, the modelling method was not described, for 13% the time horizon (eg, 10 years) for which the model was predicting was not described, and for 25% information for calculating individual CVD risks (eg, full regression equation, nomogram, or risk chart) was insufficient, making it impossible to validate these models or apply them in clinical practice. For external validation of a model, the full regression equation is needed, which was presented for only 46% of the developed models. To improve the reporting of prediction model studies, the TRIPOD statement was recently published (www.tripod-statement.org).[30,57]

Since the publication of the review by Beswick et al[16] in 2008, in which they searched the literature until 2003, several major things have changed. The number of developed prediction models has more than tripled, from 110 to 363, revealing problems such as the overwhelming number of prediction models, predictor definitions, outcome definitions, prediction horizons, and study populations, and showing how poorly researchers make use of available evidence or existing models in the discipline. Although Beswick et al stated that -New prediction models should have multiple external validations in diverse populations with differing age ranges, ethnicity, sex and cardiovascular risk-,[16] we still found a great lack of validation studies for most developed CVD risk prediction models.

Presumably there are various reasons why researchers continue to develop a new CVD risk prediction model from scratch, such as the perceived lack of prediction models for their specific population (eg, ethnic minority groups) or specific outcomes (eg, ischaemic stroke), newly identified predictors, published articles reporting on bad performance of existing models in another setting, availability of data with higher quality (eg, greater sample size, prospectively collected data), funding priorities, or merely self-serving to generate another publication. Nevertheless, our review clearly indicates that many of these studies are still similar in design and execution, as corresponding models often include the same (or similar) predictors, target the same (or similar) patient populations, and predict the same (or similar) outcomes. Therefore, researchers are often—perhaps without knowing—repeating the same process and mostly introduce implicit knowledge when developing a prediction model from scratch. Given that there is a huge amount of literature on prediction of CVD outcomes for the general population, we think it is time to capitalise on prediction modelling research from scratch in this specialty. Over the past few decades, statistical methods for building prediction models using established knowledge have substantially improved, and these can be achieved by refining, updating, extending, and even combining the most promising existing models for prediction of CVD in the general population.

Recommendations and policy implications

Ideally, systematic reviews also guide evidence informed health decision making, in this case leading to recommendations on which models to advocate or even use in different settings or countries. Given the lack of external validation studies (notably by independent investigators) of the majority of CVD risk prediction models, the even bigger lack of head-to-head comparisons of these models (even of the well known CVD risk prediction models such as Framingham, SCORE, and QRISK), the poor reporting of most developed models, and the large variability in studied populations, predicted outcomes, time horizons, included predictors, and reported performance measures, we believe it is still impossible to recommend which specific model or models should be used in which setting or location. Guided by this review, we will continue to focus on quantitatively summarising the predictive performance of the identified CVD risk prediction models that were externally validated across various different locations, and ideally of models that were validated head-to-head and compared in the same dataset. Such meta-analysis of CVD risk prediction models should attempt to identify boundaries of the external validity and thus eventual applicability of these frequently validated models.

This leads to a number of new recommendations in the discipline of CVD risk prediction research and practice. Firstly, this area would benefit from the formulation of guidance with clear definitions of the relevant outcomes (eg, similar to the CROWN initiative in obstetrics[58]), predictors, and prediction horizons. Secondly, the validity, and thus potential impact, of cardiovascular risk prediction models could substantially be improved by making better use of existing evidence, rather than starting from scratch to develop yet another model.[59] Thirdly, the suitable and promising models for a particular targeted population, outcome, and prediction horizon, should be identified, and subsequently be validated (and if necessary tailored to the situation at hand), allowing for head-to-head comparisons such as previously done for prediction models for type 2 diabetes[60] and patients requiring cardiac surgery.[61] Fourthly, more work is needed to evaluate the presence of heterogeneity in performance of different models across countries, allowing for tailoring of prediction models to different subpopulations. This can be achieved by combining the individual participant data (IPD) from multiple sources, including the increasingly available large registry datasets, and performing the so called IPD meta-analysis.[62,63] Analysis of such combined or large datasets has the advantage not only of increased total sample size, but also of better tackling case mix effects, setting specific issues (eg, inclusion of setting specific predictors), and better tailoring of existing models to different settings and consequently improving the robustness and thus generalisability of prediction models across subgroups and countries. Recently, prediction modelling methods for analysis of large, combined datasets have been proposed.[59,63-68] If, after these efforts, generalisability of a developed and validated prediction model is still not good enough (eg, because of too much differences between populations, treatment standards, or data quality), more advanced

methods for redevelopment of models can be used. Promising techniques are dynamic prediction modelling,[69,70] modelling strategies that take into account treatment-covariate interactions,[71] or other techniques such as machine learning.[72,73] Finally, models with adequate generalisability -as inferred from external validation studies- should be evaluated for potential impact on doctors' decision making or patient outcomes, before being incorporated in guidelines.[20,74] A recently published systematic review showed that the provision of risk information increases prescribing of antihypertensive drugs and lipid lowering drugs, but to our knowledge there are yet no studies investigating the effect of the use of prediction models and risk information provision on actual incidences of CVD events.[27]

Conclusions

The current literature is overwhelmed with models for predicting the risk of cardiovascular outcomes in the general population. Most, however, have not been externally validated or directly compared on their relative predictive performance, making them currently of yet unknown value for practitioners, policy makers, and guideline developers. Moreover, most developed prediction models are insufficiently reported to allow external validation by others, let alone to become implemented in clinical guidelines or being used in practice. We believe it is time to stop developing yet another similar CVD risk prediction model for the general population. Rather than developing such new CVD risk prediction models, in this era of large and combined datasets, we should focus on externally validating and comparing head-to-head the promising existing CVD risk models, on tailoring these models to local settings, to investigate whether they may be extended with new predictors, and finally to quantify the clinical impact of the most promising models.

Acknowledgments

We thank René Spijker for performing the literature search and Johannes B Reitsma who provided insight and expertise that greatly assisted this project.

References

1. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol* 2013;13:33.
2. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ* 2009;338:b375.
3. Eckel RH, Jakicic JM, Ard JD, de Jesus JM, Houston Miller N, Hubbard VS, et al. 2013 AHA/ACC guideline on lifestyle management to reduce cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014;129(25 Suppl 2):S76-99.
4. Alwan A. *Global status report on noncommunicable diseases 2010*: World Health Organization, 2011.
5. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. *Am Heart J* 1991;121(1 Pt 2):293-8.
6. Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk profile. A statement for health professionals. *Circulation* 1991;83(1):356-62.
7. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97(18):1837-47.
8. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J* 2003;24(11):987-1003.
9. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *BMJ* 2010;341:c6624.
10. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007;335(7611):136.
11. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336(7659):1475-82.
12. Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014;129(25 Suppl 2):S49-73.
13. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11(10):e1001744.

14. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15(4):361-87.
15. Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. NICE guidelines [CG181], National Institute for Health and Clinical Excellence 2014.
16. Beswick AD, Brindle P, Fahey T, Ebrahim S. *A Systematic Review of Risk Scoring Methods and Clinical Decision Aids Used in the Primary Prevention of Coronary Heart Disease (Supplement)*. London: Royal College of General Practitioners, 2008.
17. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23(16):2567-86.
18. Matheny M, McPheeters ML, Glasser A, Mercaldo N, Weaver RB, Jerome RN, et al. *Systematic Review of Cardiovascular Disease Risk Assessment Tools*. Rockville MD, 2011.
19. Wessler BS, Lai Yh L, Kramer W, Cangelosi M, Raman G, Lutz JS, et al. Clinical Prediction Models for Cardiovascular Disease: Tufts Predictive Analytics and Comparative Effectiveness Clinical Prediction Model Database. *Circ Cardiovasc Qual Outcomes* 2015;8(4):368-75.
20. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009;338:b606.
21. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012;98(9):683-90.
22. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ* 2009;338:b604.
23. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;35(29):1925-31.
24. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19(4):453-73.
25. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:b605.
26. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Invest* 2012;42(2):216-28.
27. Usher-Smith JA, Silarova B, Schuit E, Gm Moons K, Griffin SJ. Impact of provision of cardiovascular disease risk estimates to healthcare professionals and patients: a systematic review. *BMJ Open* 2015;5(10):e008717.

28. van Dieren S, Beulens JW, Kengne AP, Peelen LM, Rutten GE, Woodward M, et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. *Heart* 2012;98(5):360-9.
29. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9(5):1-12.
30. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162(1):55-63.
31. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. *BMJ* 2012;344:e3318.
32. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10(2):e1001381.
33. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol* 1995;48(12):1503-10.
34. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49(12):1373-9.
35. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21(1):128-38.
36. D'Agostino RB, Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008;117(6):743-53.
37. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* 2002;106(25):3143-421.
38. Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. *Circulation* 2002;105(3):310-5.
39. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham Study. *Stroke* 1991;22(3):312-8.
40. Chambless LE, Folsom AR, Sharrett AR, Sorlie P, Couper D, Szklo M, et al. Coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC) study. *J Clin Epidemiol* 2003;56(9):880-90.

41. Friedland DR, Cederberg C, Tarima S. Audiometric pattern as a predictor of cardiovascular status: development of a model for assessment of risk. *Laryngoscope* 2009;119(3):473-86.
42. Keys A, Aravanis C, Blackburn H, Van Buchem FS, Buzina R, Djordjevic BS, et al. Probability of middle-aged men developing coronary heart disease in five years. *Circulation* 1972;45(4):815-28.
43. Leaverton PE, Sorlie PD, Kleinman JC, Dannenberg AL, Ingster-Moore L, Kannel WB, et al. Representativeness of the Framingham risk model for coronary heart disease mortality: a comparison with a national cohort study. *J Chronic Dis* 1987;40(8):775-84.
44. Barzi F, Patel A, Gu D, Sritara P, Lam TH, Rodgers A, et al. Cardiovascular risk prediction tools for populations in Asia. *J Epidemiol Community Health* 2007;61(2):115-21.
45. Woodward M, Brindle P, Tunstall-Pedoe H. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart* 2007;93(2):172-6.
46. Levy D, Wilson PW, Anderson KM, Castelli WP. Stratifying the patient at risk from coronary disease: new insights from the Framingham Heart Study. *Am Heart J* 1990;119(3 Pt 2):712-7; discussion 17.
47. Chien KL, Hsu HC, Su TC, Chang WT, Chen PC, Sung FC, et al. Constructing a point-based prediction model for the risk of coronary artery disease in a Chinese community: a report from a cohort study in Taiwan. *Int J Cardiol* 2012;157(2):263-8.
48. Collins GS, Moons KG. Comparing risk prediction models. *BMJ* 2012;344:e3186.
49. WHO. Cardiovascular diseases (CVDs) Fact sheet N°317, 2016.
50. Hajifathalian K, Ueda P, Lu Y, Woodward M, Ahmadvand A, Aguilar-Salinas CA, et al. A novel risk score to predict cardiovascular disease risk in national populations (Globorisk): a pooled analysis of prospective cohorts and health examination surveys. *Lancet Diabetes Endocrinol* 2015;3(5):339-55.
51. Liu J, Hong Y, D'Agostino RB, Sr., Wu Z, Wang W, Sun J, et al. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study. *JAMA* 2004;291(21):2591-9.
52. Wu Y, Liu X, Li X, Li Y, Zhao L, Chen Z, et al. Estimation of 10-year risk of fatal and nonfatal ischemic cardiovascular diseases in Chinese adults. *Circulation* 2006;114(21):2217-25.
53. Gijsberts CM, Groenewegen KA, Hoefer IE, Eijkemans MJ, Asselbergs FW, Anderson TJ, et al. Race/Ethnic Differences in the Associations of the Framingham Risk Factors with Carotid IMT and Cardiovascular Events. *PLoS One* 2015;10(7):e0132321.
54. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med* 2011;9:103.

55. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013;66(3):268-77.
56. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 2014;383(9913):267-76.
57. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.
58. Khan KS, Romero R. The CROWN initiative: journal editors invite researchers to develop core outcomes in women's health. *Am J Obstet Gynecol* 2014;211(6):575-6.
59. Debray TP, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW, Moons KG. Meta-analysis and aggregation of multiple published prediction models. *Stat Med* 2014;33(14):2341-62.
60. Abbasi A, Peelen LM, Corpeleijn E, van der Schouw YT, Stolk RP, Spijkerman AM, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ* 2012;345:e5900.
61. Ettema RG, Peelen LM, Schuurmans MJ, Nierich AP, Kalkman CJ, Moons KG. Prediction models for prolonged intensive care unit stay after cardiac surgery: systematic review and validation study. *Circulation* 2010;122(7):682-9, 7 p following p 89.
62. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2014.
63. Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS Med* 2015;12(10):e1001886.
64. Debray TP, Koffijberg H, Vergouwe Y, Moons KG, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. *Stat Med* 2012;31(23):2697-712.
65. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* 2013;32(18):3158-80.
66. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2015.
67. Royston P, Parmar MK, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Stat Med* 2004;23(6):907-26.

68. Sauerbrei W, Royston P. A new strategy for meta-analysis of continuous covariates in observational studies. *Stat Med* 2011;30(28):3341-60.
69. Nicolaie MA, van Houwelingen JC, de Witte TM, Putter H. Dynamic prediction by landmarking in competing risks. *Stat Med* 2013;32(12):2031-47.
70. Teramukai S, Okuda Y, Miyazaki S, Kawamori R, Shirayama M, Teramoto T. Dynamic prediction model and risk assessment chart for cardiovascular disease based on on-treatment blood pressure and baseline risk factors. *Hypertens Res* 2016;39(2):113-8.
71. van Klaveren D, Vergouwe Y, Farooq V, Serruys PW, Steyerberg EW. Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. *J Clin Epidemiol* 2015;68(11):1366-74.
72. Wolfson J, Bandyopadhyay S, Elidrisi M, Vazquez-Benitez G, Vock DM, Musgrove D, et al. A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data. *Stat Med* 2015;34(21):2941-57.
73. Guo Y, Wei Z, Keating BJ, Hakonarson H. Machine learning derived risk prediction of anorexia nervosa. *BMC Med Genomics* 2016;9(1):4.
74. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;98(9):691-8.

Supplemental material

Supplemental Table 1: Medline search strategy.

1	chd risk assessment$.mp.
2	cvd risk assessment$.mp.
3	heart disease risk assessment$.mp.
4	coronary disease risk assessment$.mp.
5	cardiovascular disease risk assessment$.mp.
6	cardiovascular risk assessment$.mp.
7	cv risk assessment$.mp.
8	cardiovascular disease$ risk assessment$.mp.
9	coronary risk assessment$.mp.
10	coronary risk scor$.mp.
11	heart disease risk scor$.mp.
12	chd risk scor$.mp.
13	cardiovascular risk scor$.mp.
14	cardiovascular disease$ risk scor$.mp.
15	cvd risk scor$.mp.
16	cv risk scor$.mp.
17	or/1-16
18	cardiovascular diseases/
19	coronary disease/
20	cardiovascular disease$.mp.
21	heart disease$.mp.
22	coronary disease$.mp.
23	cardiovascular risk?.mp.
24	coronary risk?.mp.
25	exp hypertension/
26	exp hyperlipidemia/
27	or/18-26
28	risk function.mp.
29	Risk Assessment/mt
30	risk functions.mp.
31	risk equation$.mp.
32	risk chart?.mp.
33	(risk adj3 tool$).mp.
34	risk assessment function?.mp.
35	risk assessor.mp.
36	risk appraisal$.mp.
37	risk calculation$.mp.
38	risk calculator$.mp.
39	risk factor$ calculator$.mp.

40 risk factor$ calculation$.mp.
41 risk engine$.mp.
42 risk equation$.mp.
43 risk table$.mp.
44 risk threshold$.mp.
45 risk disc?.mp.
46 risk disk?.mp.
47 risk scoring method?.mp.
48 scoring scheme?.mp.
49 risk scoring system?.mp.
50 risk prediction?.mp.
51 predictive instrument?.mp.
52 project$ risk?.mp.
53 cdss.mp.
54 or/28-53
55 27 and 54
56 17 or 55
57 new zealand chart$.mp.
58 sheffield table$.mp.
59 procam.mp.
60 General Rule to Enable Atheroma Treatment.mp.
61 dundee guideline$.mp.
62 shaper scor$.mp.
63 (brhs adj3 score$).mp.
64 (brhs adj3 risk$).mp.
65 copenhagen risk.mp.
66 precard.mp.
67 (framingham adj1 (function or functions)).mp.
68 (framingham adj2 risk).mp.
69 framingham equation.mp.
70 framingham model$.mp.
71 (busselton adj2 risk$).mp.
72 (busselton adj2 score$).mp.
73 erica risk score$.mp.
74 framingham scor$.mp.
75 dundee scor$.mp.
76 brhs scor$.mp.
77 British Regional Heart study risk scor$.mp.
78 brhs risk scor$.mp.
79 dundee risk scor$.mp.

80 framingham guideline$.mp.
81 framingham risk?.mp.
82 new zealand table$.mp.
83 ncep guideline?.mp.
84 smac guideline?.mp.
85 copenhagen risk?.mp.
86 or/57-85
87 56 or 86
88 exp decision support techniques/
89 Diagnosis, Computer-Assisted/
90 Decision Support Systems,Clinical/
91 algorithms/
92 algorithm?.mp.
93 algorythm?.mp.
94 decision support?.mp.
95 predictive model?.mp.
96 treatment decision?.mp.
97 scoring method$.mp.
98 (prediction$ adj3 method$).mp.
99 or/88-98
100 Risk Factors/
101 exp Risk Assessment/
102 (risk? adj1 assess$).mp.
103 risk factor?.mp.
104 or/100-103
105 27 and 99 and 104
106 87 or 105
107 stroke.mp.
108 exp Stroke/
109 cerebrovascular.mp. or exp Cerebrovascular Circulation/
110 limit 106 to ed=20040101-20130601
111 107 or 108 or 109
112 111 and 54
113 111 and 99 and 104
114 112 or 113
115 106 or 114

Supplemental Table 2: List of articles in which the development of a model was presented, the number of models that were developed in these articles and references of papers in which these models were externally validated or incremental value was assessed.

First author, publication year	Number of models developed	Number of articles in which model is validated
Adult Treatment Panel III 2002[1]	2	19[2-20]
Alssema 2012[21]	2	-
Anderson 1991a[22]	12	28[23-50]
Anderson 1991b[51]	4	10[9,25,52-59]
Arima 2009[60]	1	-
Asayama 2008[61]	2	-
Asia Pacific Cohort Studies Collaboration 2006[62]	2	-
Asia Pacific Cohort Studies Collaboration 2007[63]	4	1[63]
Aslibekyan 2011[64]	2	1[64]
Assmann 2002[65]	1	6[5,7,66-69]
Assmann 2007[70]	3	1[71]
Assmann 2008[66]	1	-
Balkau 2004[72]	8	-
Bastuji 2002[23]	6	-
Beer 2011[73]	1	-
Bell 2012[74]	4	-
Berard 2011[75]	1	-
Boland 2004[76]	1	1[76]
Bolton 2013[77]	1	-
Boudik 2006[52]	1	-
Brand 1976[78]	1	-
Braun 2013[79]	6	1[79]
Brautbar 2009[80]	2	-
Brindle 2006[81]	32	-
Chamberlain 2011[3]	2	-
Chambless 2003[82]	4	3[83-85]
Chen 2009[27]	2	1[27]
Chien 2010[71]	2	-
Chien 2012[67]	3	1[67]

Supplemental Table 2: Continued

First author, publication year	Number of models developed	Number of articles in which model is validated
Ciampi 2001[86]	10	-
Conroy 2003[87]	12	21[24,27,43,44,55,85,88-102]
Cook 2006[4]	1	-
Cooper 2005[5]	1	-
Cross 2012[103]	1	1[103]
D'Agostino 1994[104]	2	1[105]
D'Agostino 2000[106]	2	-
D'Agostino 2008[107]	4	15[6,20,28,43,50,53,73,108-115]
Davies 2010[116]	1	-
De Ruijter 2009[31]	5	-
Donfrancesco 2010[117]	2	-
Dunder 2004[7]	1	-
Duprez 2011[8]	1	-
Empana 2011[118]	1	-
Faeh 2013[119]	2	-
Ferrario 2005[69]	1	-
Folsom 2003[120]	1	1[3]
Friedland 2009[121]	7	1[121]
Gaziano 2008[122]	4	1[43]
Glynn 2002[123]	2	-
Hesse 2005[124]	1	1[124]
Hippisley-Cox 2007[34]	2	4[28,29,33,35]
Hippisley-Cox 2008b[35]	2	1[30]
Hippisley-Cox 2010[125]	2	1[30]
Hoes 1993[126]	2	-
Houterman 2002[127]	2	-
Ishikawa 2009[128]	3	-
Janssen 2005[129]	1	-
Kannel 1976[130]	2	-
Keys 1972[131]	4	1[131]
Knuiman 1997[132]	4	-
Knuiman 1998[133]	2	-

Supplemental Table 2: Continued

First author, publication year	Number of models developed	Number of articles in which model is validated
Koller 2012[10]	4	-
L'Italien 2000[134]	1	1[134]
Larson 1995[135]	3	-
Leaverton 1987[136]	4	2[132,136]
Lee 2006[137]	2	-
Lee 2008[138]	4	-
Levy 1990[139]	4	1[132]
Liu 2004[140]	2	-
Lloyd-Jones 2006[141]	2	1[47]
Lumley 2002[142]	2	1[143]
Macfarlane 2007[144]	1	-
Mainous 2007[93]	3	-
Mannan 2010[145]	2	-
Mannan 2011[146]	1	-
Mannan 2013[147]	2	-
Matsumoto 2009[148]	2	-
May 2006[37]	2	-
May 2007[149]	1	-
McGorrian 2011[150]	4	1[150]
McNeil 2001[151]	1	-
Menotti 1990[152]	1	-
Menotti 1994[153]	1	1[153]
Menotti 2000[154]	3	-
Menotti 2002[155]	3	-
Menotti 2005[156]	2	-
Merry 2012[94]	1	-
Moons 2002[157]	3	-
Nelson 2012[40]	1	1[40]
Nippon Data Research Group 2006[158]	6	-
Noda 2010[159]	3	-
Nordestgaard 2010[160]	1	-
Odell 1994[161]	9	-

Supplemental Table 2: Continued

First author, publication year	Number of models developed	Number of articles in which model is validated
Onat 2012[162]	2	-
Panagiotakos 2007[96]	2	-
Pencina 2009[163]	1	-
Petersson 2009[164]	2	-
Plichart 2011[165]	2	-
Pocock 2001[166]	1	1[41]
Polonsky 2010[167]	2	-
Prati 2011[168]	1	-
Qiao 2012[169]	8	-
Ridker 2007[15]	2	1[20]
Ridker 2008[16]	4	1[20]
Schnabel 2009[170]	1	1[3]
Shaper 1986[171]	2	-
Simons 2003[18]	2	1[41]
Smith 2010[172]	2	-
Tanabe 2010[173]	2	-
Teramoto 2008[174]	1	-
Thomsen 2001[175]	1	-
Thorsen 1979[176]	1	-
Truett 1967[177]	2	1[78]
Tsang 2003[178]	1	1[178]
Tunstall-Pedoe 1991[179]	2	1[179]
Vergnaud 2008[180]	1	-
Voss 2002[181]	2	1[182]
Wilson 1987[183]	2	1[67]
Wilson 1998[184]	2	41[2,3,15,54,68,69,88-90,93-95,98,101,103,107,138,140,180,182,185-205]
Wolf 1991[206]	2	5[71,142,143,207,208]
Woodward 2007[48]	2	2[34,53]
Wu 2006[209]	2	1[209]
Wu 2011[210]	2	-
Yip 2004[211]	1	-

Supplemental Table 2: Continued

First author, publication year	Number of models developed	Number of articles in which model is validated
Zhang 2005[49]	3	-
Framingham unspecified*	-	3[84,85,212]

*If authors explicitly stated they determined incremental value on top of the variables from a Framingham model without referencing this specific model, they were categorized under Framingham unspecified.

Supplemental Table 3: Main categories of outcomes that were used in the developed models.

Outcome	N (%)
Fatal or nonfatal CHD	118 (33%)
Fatal or nonfatal CVD	95 (26%)
Fatal CVD	40 (11%)
Fatal or nonfatal stroke	29 (8%)
Fatal or nonfatal MI	23 (6%)
Fatal CHD	21 (6%)
All-cause mortality	9 (2%)
Atrial fibrillation	4 (1%)
Fatal nonCHD	4 (1%)
Fatal or nonfatal stroke, TIA	4 (1%)
Ischemic stroke	3 (1%)
Fatal stroke	2 (1%)
Hemorrhagic stroke	2 (1%)
Nonfatal MI	2 (1%)
Claudication	1 (<0.5%)
Coronary artery bypass grafting	1 (<0.5%)
Heart failure	1 (<0.5%)
Ischemic stroke, TIA	1 (<0.5%)
Nonfatal CHD	1 (<0.5%)
Percutaneous transluminal coronary angioplasty	1 (<0.5%)
TIA	1 (<0.5%)
Total	363

CHD=coronary heart disease; CVD=cardiovascular disease; MI=myocardial infarction; TIA=transient ischemic attack.

Supplemental Table 4: Outcome definitions as extracted by reviewers, and category in which these were placed of developed models.

Outcome category	Definition
Fatal or nonfatal CHD (n=118)	Any fatal/non-fatal coronary event: death from CHD or definite myocardial infarction, and any CHD, classical angina pectoris, clinical judgment of definite heart disease and etiology specified as myocardial infarction by history, and (3) follow-up clinical diagnosis of possible heart disease with etiology specified by history as myocardial infarction and any of Minnesota ECG codes17 1.2, 1.3, 5.1, 5.2, 6.1, 6.2, 7.1, 7.2, 7.4, or 8.3 at the 5-year examination, or Minnesota ECG codes 1.2 or 1.3 + 5.1 or 1.3 + 5.2 at the 5-year examination but not at entry.
	CHD death or hospitalization: ICD-9 410-414
	CHD event: a validate definite or probable hospitalized myocardial infarction, a definite CHD death, an unrecognised myocardial infarction defined by ARIC ECG readings, or coronary revascularization.
	CHD hard criteria: CHD death (ICD-9 410-414 or code 428.0-1), definite MI
	CHD-any criterion: CHD death (ICD-9 410-414 or code 428.0-1), fatal or non-fatal MI, angina pectoris, chronic heart disease of possible coronary origin, coronary bypass surgery, coronary angioplasty
	CHD: all definite myocardial infarction, coronary insufficiency, angina pectoris and death from coronary heart disease.
	CHD: death from CHD (sudden or non-sudden), myocardial infarction, angina pectoris and coronary insufficiency
	CHD: definite or probable myocardial infarction, silent myocardial infarction between examinations (indicated by ECG), definite CHD death, coronary revascularization
	CHD: fatal and non-fatal myocardial infarction, angina pectoris, cardiac/ sudden death, and angioplasty
	CHD: fatal and non-fatal myocardial infarction, cardiovascular death, angina pectoris
	CHD: ICD-9: 410-414
	CHD: presence of angina pectoris, a history of myocardial infarction with or without accompanying Minnesota codes of the ECG, a history of myocardial revascularisation, death from heart failure of coronary origin and fatal coronary event
	CHD: sudden coronary death, fatal acute myocardial infarction, nonfatal acute myocardial infarction, new major Q wave on the ECG after 5 years of follow-up (Minnesota codes 11, 12.1 to 12.7, and 12.8 plus 51 or 52) surgery for angina pectoris with CHD angiographically demonstrated

Supplemental Table 4: Continued

Outcome category	Definition
	CHD: validated definite or probable hospitalized MI, a definite CHD death, an unrecognized MI defined by ARIC ECG readings, or coronary revascularization. The criteria for definite or probable hospitalized MI were based on combinations of chest pain symptoms, ECG changes, and cardiac enzyme levels [33,34]. The criteria for definite fatal CHD were based on chest pain symptoms, underlying cause of death from the death certificate, and any other associated hospital information or medical history, including that from the ARIC clinic visit
	Coronary artery disease
	Coronary artery disease or coronary artery disease death (angina pectoris, myocardial infarction, coronary insufficiency)
	Coronary death, MI, angina, coronary insufficiency
	Coronary deaths, underlying causes of death ICD-IX codes 410-414, 798, 799, 250, 428, 440 in association with 410-414 codes in other causes were considered as suspected coronary deaths Non fatal coronary events: ICD IX 410–411 codes for suspected acute infarction and ICD IX CM 36.0-9 codes for coronary surgery revascularization.
	Coronary heart disease
	Coronary heart disease (MI, CHD death, angina pectoris, coronary insufficiency)
	Coronary heart disease events: myocardial infarction or death from coronary heart disease (ICD-9 codes 410-414).
	Coronary heart disease: angina pectoris, coronary insufficiency (unstable angina), myocardial infarction, and sudden death
	Coronary heart disease: angina pectoris, coronary insufficiency, myocardial infarction (recognized or not), sudden death
	Coronary heart disease: angina pectoris, recognized and unrecognized myocardial infarction, coronary insufficiency, and coronary heart disease death
	Coronary heart disease: coronary revascularization or fatal or nonfatal myocardial infarction
	Coronary heart disease: hospitalization for angina pectoris, myocardial infarction, or a CHD death (I210–I219, I251–I259, I461 and R960 ICD-10 codes), or a revascularization procedure (percutaneous intervention or coronary artery bypass- grafting).
	Coronary heart disease: MI or acute coronary death
	Coronary heart disease: MI, aorto-coronary bypass, angina, other forms of specifically defined ischemic cardiopathies or cardiac complications before or after surgery
	Coronary heart disease: myocardial infarction, death due to CHD, resuscitated cardiac arrest, definite or probable angina followed by coronary revascularization, and definite angina not followed by coronary revascularization

Supplemental Table 4: Continued

Outcome category	Definition
	Coronary mortality, non-fatal myocardial infarction
	Definite fatal coronary heart disease or definite nonfatal myocardial infarction
	Fatal and nonfatal CHD: angina pectoris and myocardial infarction (ICD-9 Codes: 410–414.9)
	Fatal or non-fatal myocardial infarction, angioplasty, coronary artery bypass surgery
	Fatal or nonfatal CHD: nonfatal definite MI, definite CHD, ECG-evident definite MI, fatal definite MI, definite CHD, possible CHD (87% fatal), and sudden death due to CHD
	First coronary heart disease event
	First major coronary event (definition reported in unavailable article)
	Hard CHD: acute myocardial infarction, sudden death, and other coronary deaths
	hard CHD: death from CHD or definite myocardial infarction, and any CHD (classical angina pectoris, (2) clinical judgment of definite heart disease and etiology specified as myocardial infarction by history, and (3) follow-up clinical diagnosis of possible heart disease with etiology specified by history as myocardial infarction and any of Minnesota ECG codes17 1.2, 1.3, 5.1, 5.2, 6.1, 6.2, 7.1, 7.2, 7.4, or 8.3 at the 5-year examination, or Minnesota ECG codes 1.2 or 1.3 + 5.1 or 1.3 + 5.2 at the 5-year examination but not at entry).
	Hard CHD: myocardial infarction + CHD death
	Incident cases of coronary heart disease: death with an underlying or contributing cause of CHD (International classification of diseases, 10 revision codes I20–I25, I51.6) or a myocardial infarction, diagnosis of angina or coronary artery bypass or angioplasty identified in the follow-up medical record review.
	Incident coronary heart disease: a clinical diagnosis of an acute myocardial infarction, unstable angina pectoris, a percutaneous transluminal coronary angioplasty, or coronary artery bypass grafting according to the Cardiology information system or coronary heart disease as primary or secondary cause of death according to Statistics Netherlands (ICD9 410–414 or ICD10 I20–I25).
	Incident coronary heart disease: fatal and nonfatal myocardial infarction, percutaneous coronary intervention, coronary artery bypass graft
	Incident coronary heart disease: myocardial infarction, fatal coronary heart disease, cardiac procedure
	Ischemic cardiovascular disease: acute myocardial infarction, coronary death, ischemic cardiac arrest, ischemic stroke (brain infarction due to occlusion of precerebral arteries or embolic brain infarction, ICD-9 433-434)
	Major coronary event: sudden cardiac death or definite fatal or nonfatal myocardial infarction on the basis of ECG and/or cardiac enzyme changes

Supplemental Table 4: Continued

Outcome category	Definition
	Major coronary event: sudden cardiac death, definite fatal or non-fatal myocardial infarction on the basis of ECG and/or cardiac enzyme changes. The detailed criteria for defining a sudden coronary death and a definite fatal or non-fatal myocardial infarction have been previously published.
	Major coronary events: nonfatal MI and coronary deaths
	Major coronary events: sudden coronary death, non-sudden coronary death, definite non-fatal myocardial infarction, fatal myocardial infarction, definite fatal chronic ischemic heart disease, surgery of coronary arteries
	Myocardial infarction, undergone coronary artery bypass grafting, had percutaneous coronary intervention, or had a coronary angiography or computed tomography angiography demonstrating a stenosis of at least 50% in at least 1 epicardial vessel
	Non-fatal myocardial infarction (ECG and/or cardiac enzyme changes), fatal MI (MI 28 d before death and no known nonatherosclerotic cause of death), atherosclerotic CHD death (Chest pain 72 h before death and no known nonatherosclerotic cause of death; History of chronic ischemic heart disease in the absence of valvular heart disease or nonischemic cardiomyopathy and no known nonatherosclerotic cause of death; Death certificate consistent with atherosclerotic CHD death and no known nonatherosclerotic cause of death; Coronary death related to CHD procedures, such as PCI or CABG)
Fatal or nonfatal CVD (n=95)	Atherosclerotic CVD: ICD-8 D410-D414, D427, D430-438, D440-444
	Cardiovascular disease
	Cardiovascular disease event: coronary heart disease (CHD) and stroke
	Cardiovascular disease event: coronary heart disease (CHD) and stroke - Definition: referred to 5 references with different definitions
	Cardiovascular disease: coronary heart disease (angina and myocardial infarction), stroke, or transient ischaemic attacks in the term cardiovascular disease but not peripheral vascular disease. ICD-10 codes I20-I25, I63-I64.
	Cardiovascular disease: coronary heart disease, congestive heart failure, cerebrovascular disease, intermittent claudication
	Cardiovascular disease: includes coronary heart disease (angina and myocardial infarction), stroke, or transient ischaemic attacks, but not peripheral vascular disease.
	Cardiovascular disease: myocardial infarction, coronary heart disease, stroke, and transient ischaemic attack.
	Cardiovascular disease: myocardial infarction, coronary insufficiency, death resulting from coronary heart disease, angina pectoris, atherothrombotic stroke, intermittent claudication, or other cardiovascular death

Supplemental Table 4: Continued

Outcome category	Definition
	Cardiovascular disease: stroke or coronary heart disease including acute myocardial infarction, silent myocardial infarction, sudden cardiac death within 1 h after onset of acute illness, coronary artery disease followed by coronary artery bypass surgery or angioplasty. Cardiovascular disease was defined as first-ever development of coronary heart disease or stroke. The criteria for a diagnosis of coronary heart disease included first-ever acute myocardial infarction, silent myocardial infarction, sudden cardiac death within 1 h after the onset of acute illness, or coronary artery disease followed by coronary artery bypass surgery or angioplasty. Acute myocardial infarction was diagnosed when a subject met at least two of the following criteria: (1) typical symptoms, including prolonged severe anterior chest pain; (2) abnormal cardiac enzymes more than twice the upper limit of the normal range; (3) evolving diagnostic electrocardiographic changes; and (4) morphological changes, including local asynergy of cardiac wall motion on echocardiography, persistent perfusion defect on cardiac scintigraphy, or myocardial necrosis or scars 41 cm long accompanied by coronary atherosclerosis at autopsy. Silent myocardial infarction was defined as myocardial scarring without any historical indication of clinical symptoms or abnormal cardiac enzyme changes, and was detected by electrocardiography, echocardiography, cardiac scintigraphy or autopsy. Stroke was defined as a sudden onset of nonconvulsive and focal neurological deficit persisting for 424 h. The diagnosis of stroke and the determination of its pathological type were based on the clinical history, neurological examination and all available clinical data, including brain CT/MRI and autopsy findings.
	Cardiovascular events (myocardial infarction, ischemic stroke, coronary revascularization, cardiovascular death).
	CHD, ischemic stroke and MI (ICD codes of 433–434 (I63), 410–414 (I20–I25) and 410–411 (I21–I22, I24))
	CHF, AF, MI, coronary revascularisation, stroke, transient ischemic attack and CVD death.
	CVD including coronary heart disease, stroke, or peripheral vascular disease
	CVD-any criterion: CHD death (ICD-9 410-414 or code 428.0-1), fatal or non-fatal MI, angina pectoris, chronic heart disease of possible coronary origin, coronary bypass surgery, coronary angioplasty, cerebrovascular death (ICD-9 430-438), stroke, TIA, peripheral artery disease, intermitted claudication, aortic aneurysm, arterial surgical procedures
	CVD: cardiovascular death, non-fatal myocardial infarction or non-fatal cerebrovascular event
	CVD: CHD (coronary death, myocardial infarction, coronary insufficiency, and angina), cerebrovascular events (including ischaemic stroke, haemorrhagic stroke, and transient ischaemic attack), peripheral artery disease (intermittent claudication), and heart failure
	CVD: CHD (coronary death, myocardial infarction, coronary insufficiency, and angina), cerebrovascular events (ischemic stroke, haemorrhagic stroke, transient ischemic attack), peripheral artery disease (intermittent claudication), heart failure.

Supplemental Table 4: Continued

Outcome category	Definition
	CVD: coronary heart disease (angina pectoris, coronary insufficiency, myocardial infarction, sudden or non-sudden death attributed to coronary disease), cerebrovascular accident (stroke, transient ischaemia, cerebral embolism, intracerebral or subarachnoid haemorrhage), intermittent claudication, and congestive heart failure
	CVD: death from CHD (sudden or non-sudden), myocardial infarction, angina pectoris, coronary insufficiency, stroke and transient ischemia
	CVD: death, myocardial infarction, stroke, congestive heart failure, and coronary revascularisation including coronary artery bypass grafting and percutaneous transluminal coronary angioplasty
	CVD: fatal and non-fatal myocardial infarction, cardiovascular death, angina pectoris, fatal and non-fatal stroke, transient ischaemic attack and subarachnoid haemorrhage, fatal and non-fatal heart failure and cerebrovascular death of other origin
	CVD: MI, CHD death, angina pectoris, coronary insufficiency, stroke, congestive heart failure. peripheral vascular disease
	CVD: myocardial infarction, angina, stroke, coronary artery bypass surgery, percutaneous coronary intervention, heart failure, peripheral vascular disease
	CVD: myocardial infarction, angina, stroke, left ventricular or congestive cardiac failure, peripheral vascular event, sudden/rapid cardiac death, heart failure death or other coronary or cardiovascular death
	CVD: myocardial infarction, coronary death or stroke. This outcome (effectively "hard" CHD) excluded other, non-fatal forms of CHD, but included transient ischaemic attack.
	CVD: myocardial infarction, ischemic stroke, coronary revascularization procedures, deaths from cardiovascular causes
	Deaths from cardiovascular causes (ICD-9 codes 390–459, ICD-10 codes I00-I99) or any hospital discharge diagnosis post recruitment (potentially several per admission) for coronary heart disease (ICD-9 410–414, ICD-10 I20-I25) or cerebrovascular disease (ICD-9 430–438, ICD-10 G45, I60-I69),10 11 or for coronary artery interventions (CABG or PTCA).
	Fatal and non-fatal stroke, fatal and non-fatal myocardial infarction. The International Classification of Disease codes for stroke and TIA were 362.3, 430, 431, 433.x1, 434.x1, 435, 436, G45, H34.1, I60, I61, and I6-7 and for MI were 410, 411, and I21x.
	Fatal or nonfatal CVD (myocardial infarction, percutaneous transluminal coronary angioplasty, coronary artery bypass graft, angina pectoris, stroke, claudication intermittent, peripheral intervention, or heart failure), sudden death, type 2 diabetes, and/or CKD. Cardiovascular death was defined as death due to diseases of the cardiovascular system (ICD-10: I00–I99) and sudden death (ICD-10: R96). CKD was defined by estimated glomerular filtration rate,60 mL/min/1.73 m2
	Fatal/nonfatal cardiovascular events: ICD-8 and ICD-9: 410–414, 431, 433, 434, 435, 436, 437, 440, 441, ICD-10: I20–I25, I61, I63-I66, I70–I72

Supplemental Table 4: Continued

Outcome category	Definition
	First major cardiovascular event: hard coronary event (definition reported in unavailable article), hard cerebrovascular event (definition reported in unavailable article), major peripheral artery disease (manifested as fatal peripheral artery disease, or as fatal or non-fatal gangrene of the extremities, or as fatal or non-fatal aneurysm of the aorta in any anatomical site, or as surgical procedures for aortic aneurysm or for lower limb artery disease, or as any other fatal cardiovascular event attributed to arteriosclerosis)
	First occurrence of cardiovascular disease: myocardial infarction, stroke, death from cardiovascular causes, percutaneous transluminal coronary angioplasty, or coronary artery bypass graft surgery.
	Hard cardiovascular disease: recognized MI, sudden death, or atherothrombotic brain infarction
	Hard CV events: coronary death, myocardial infarction, stroke
	Incident cardiovascular disease (CHD or stroke): death with an underlying or contribution cause coded as I20–I25, I51.6,I60–I69 or G45 or a new CHD or stroke event in any woman's medical record review.
	Major cardiovascular events: major coronary events (sudden coronary death, non-sudden coronary death, definite non-fatal myocardial infarction, fatal myocardial infarction, definite fatal chronic ischemic heart disease, surgery of coronary arteries) and cerebrovascular events (definite fatal and non-fatal haemorrhagic and thrombotic stroke, surgery of carotid arteries), plus major peripheral artery events comprising fatal and non-fatal aortic aneurysms, fatal lower limbs artery disease, surgery of aorta or lower limb arteries.
	Myocardial infarction (recognized or unrecognized), coronary heart disease, and cardiovascular disease. Specification in reference.
	Myocardial infarction, stroke, coronary revascularization procedures, or cardiovascular death
	Recognized myocardial infarction (MI) or atherothrombotic brain infarction (ABI)
Fatal CVD (n=40)	Cardiovascular death: ICD-10 codes I00-I99
	Cardiovascular death: ICD-9 codes 401–414 and 426–443, with the exception of the 430.0. 798.1 and 798.2. Instantaneous death (ICD-9, 798.1) and death within 24 h of symptoms onset (ICD-9, 798.2)
	Cardiovascular disease mortality: myocardial infarction (definite), angina pectoris (definite), intermittent claudication (definite), stroke (definite), TIA (definite) or heart failure
	Cardiovascular mortality (ICD-10: I10 to I79)
	Cardiovascular mortality: ICD-9 codes 401 through 414 and 426 through 443, with the exception of the following ICD-9 codes for definitely non-atherosclerotic causes of death: 426.7, 429.0, 430.0, 432.1, 437.3, 437.4, and 437.5. We also classified 798.1 (instantaneous death) and 798.2 (death within 24h of symptom onset) as cardiovascular deaths.
	Cause-specific death from all CVD

Supplemental Table 4: Continued

Outcome category	Definition
	CVD death: death from MI, CHD death, angina pectoris, coronary insufficiency, stroke, congestive heart failure, peripheral vascular disease
	Fatal cardiovascular events: deaths with an underlying cause given as ICD-10 codes I10 through I15, I20 through I25, R96.0, R96.1 and I44 through I73, with the exception of I45.6, I51.4, I52, I60, I62, I67.1, I67.5 and I67.7
	Fatal CVD
	Fatal CVD event: ICD-8:390–458, until 1994; ICD-10: I00-I99, since 1995
	Fatal CVD: all deaths due to ischaemic heart disease (ICD-9 410–414) and cerebrovascular accidents (ICD-9 430–438)
	Fatal CVD: ICD-8: 390–458, ICD-10: I00-I99
	Sudden death
Fatal or nonfatal stroke (n=29)	fatal/non-fatal stroke of all types
	fatal/non-fatal stroke: Atherothrombotic brain infarction, Transient ischemic attack, Cerebral embolus, Intracerebral haemorrhage, Subarachnoid haemorrhage
	First major cerebrovascular event (definition reported in unavailable article)
	Major cerebrovascular events: definite fatal and non-fatal haemorrhagic and thrombotic stroke, surgery of carotid arteries
	nonfatal ischemic stroke, transient ischemic attack (TIA) or all-causes vascular death
	Stroke
	Stroke: a clinical event of rapid onset consisting of neurological deficit lasting more than 24 hours unless death supervenes, or if it lasts less than 24 hours, an appropriate lesion to explain the deficit is seen in a brain image. The event could not be directly caused by trauma to the brain, tumour, or infection. Based on the information present, the neurologist classified the event into first and recurrent stroke, and into subarachnoid haemorrhage, intracranial haemorrhage, intracerebral infarction, or unspecified stroke. Cerebral infarction was classified according to internationally accepted criteria.22 23 In addition, the certainty of the diagnosis was assessed in definite, probable, possible and no stroke. The present analysis is restricted to definite and probable strokes
	Stroke: a focal, nonconvulsive neurological deficit of sudden onset that persisted for at least 24 hours. Stroke subtypes, ie, cerebral haemorrhage (CH), cerebral infarction (CI), and subarachnoid hemorrhage (SAH), were determined by using the criteria of the National Institute of Neurological Disorder and Stroke.23 Symptomatic lacuna infarction was defined as a CI.
	Stroke: a sudden neurological symptom of vascular origin that lasted 24 hours with supporting evidence from the image study; fatal stroke cases were included. Transient ischemic attacks were not included in this study.
	Stroke: ICD-9-CM, 430-437, or ICD-10 I60-I69
	Stroke: including transient ischemia

Supplemental Table 4: Continued

Outcome category	Definition
	Stroke: subarachnoid haemorrhage or a neurological deficit of rapid onset lasting more than 24 hours unless death supervenes or, if less than 24 hours, an appropriate lesion to explain the deficit was seen on brain imaging
Fatal or nonfatal MI (n=23)	Acute MI
	Acute MI: based on chest pain, cardiac enzyme levels, and electrocardiograms. These criteria were based on criteria from the MONICA study 28 or from the World Health Organization
	Fatal or non-fatal major ischaemic heart disease: A fatal case was considered to have occurred if ischaemic heart disease (ICD codes 410-414) was recorded as the underlying cause of death. In non-fatal cases a myocardial infarction was diagnosed according to World Health Organisation criteria
	Fatal or nonfatal MI
	Fatal or nonfatal MI: Fatal myocardial infarction was defined as cause of death with ICD-8 code 410 in the Danish National Register of Causes of Death. The nonfatal myocardial infarctions were defined as first-ever hospital admission with ICD-8 code 410 in the National Patient Register
	Fatal or nonfatal MI: Myocardial infarction was classified as bdefiniteQ or bsuspect,Q but within these categories, further subdivisions of recognized, unrecognized, and silent were made. Truly silent MI was diagnosed based on definite ECG changes (new Minnesota Code 1) without any supporting clinical history; unrecognized MI was diagnosed based on electrocardiographic changes accompanied by symptoms, which, in retrospect, were consistent with acute MI but which had not been recognized as such at the time by either the patient or his general practitioner. The diagnosis of recognized MI was based on clinical data with or without accompanying electrocardiographic abnormalities.
	Heart attack: recognized MI or sudden death
	MI
	MI (ICD 410/I21)
	MI: WHO; International Classification of Diseases, 8th edition: codes 410; 10th edition: codes I21- I22
	Myocardial infarction case: criteria from MONICA project
	Myocardial infarction: including silent and unrecognized MI
	Nonfatal or fatal definite myocardial infarction or possible myocardial infarction according to the criteria of the World Health Organization Multinational Monitoring of Trends and Determinants in Cardiovascular Disease (MONICA) Project
	Recognized myocardial infarction (MI)
Fatal CHD (n=21)	Cause-specific death from CHD
	CHD death: ICD-9 410-414
	Coronary death: ICD-9 410-414

Supplemental Table 4: Continued

Outcome category	Definition
	Coronary heart disease death (ICD-8 410, 411, 412.1 or 412.3)
	Coronary heart disease death (ICD-9 410, 411, 412 or 414)
	Coronary heart disease death: death from MI, CHD death, angina pectoris, coronary insufficiency
	Coronary heart disease mortality: death from myocardial infarction (definite) or angina pectoris (definite)
	Fatal coronary heart disease
	Fatal coronary heart disease (ICD 410-414)
	Nonsudden/sudden coronary death
All-cause mortality (n=9)	All-cause mortality
Nonfatal CHD (n=5)	Fatal nonCHD: ICD-9 codes 401 through 409 and 426 through 443, with the exception of the following ICD-9 codes for definitely non-atherosclerotic causes of death: 426.7, 429.0, 430.0, 432.1, 437.3, 437.4, and 437.5. We also classified 798.1 (instantaneous death) and 798.2 (death within 24h of symptom onset).
	Non-fatal, acute myocardial infarction (ICD-9: 410.xx and 412.xx) or hospitalization for unstable angina (ICD-9: 411.1)
Fatal or nonfatal stroke, TIA (n=4)	Fatal and non-fatal stroke, transient ischaemic attack and subarachnoid haemorrhage
	Stroke, transient ischaemic attack. The diagnostic criteria of stroke, TIA, and their subtypes were based on the system for the Classification of Cerebrovascular Disease III by the National Institute of Neurological Disorders and Stroke
Atrial fibrillation (n=4)	Atrial fibrillation: ICD-9 427.31 or 427.32
	Atrial fibrillation: 427.92 (ICD-8), 427D (ICD-9), and I48 (ICD-10)
	First event of atrial fibrillation: atrial flutter or atrial fibrillation was present on an electrocardiograph
Ischemic stroke (n=3)	Atherothrombotic brain infarction
	Ischemic stroke
	Cerebral Infarction: criteria of the National Institute of Neurological Disorder and Stroke
Fatal stroke (n=2)	Cause-specific death from stroke
Haemorrhagic stroke (n=2)	Haemorrhagic stroke
	Cerebral haemorrhage: criteria of the National Institute of Neurological Disorder and Stroke
Nonfatal MI (n=2)	Non-fatal acute myocardial infarction

Supplemental Table 4: Continued

Outcome category	Definition
Heart failure (n=1)	Heart failure: 427.00 (ICD-8), 427.10 (ICD-9), and 428.99 (ICD-10)
Ischemic stroke, TIA (n=1)	Stroke: TIA, ischaemic stroke. A transient ischaemic attack (TIA) was defined as focal neurological symptoms of ischaemic cause that lasted less than 24 h. A definite stroke was defined as a focal neurological deficit that lasted longer than 24 h and was attributable to a vascular event. Strokes were independently classified by two neurologists into ischaemic and haemorrhagic subtypes on the basis of mode of onset, clinical findings and magnetic resonance imaging and/or computerized tomography
TIA (n=1)	TIA
Claudication (n=1)	Claudication
CABG (n=1)	Coronary artery bypass grafting
PTCA (n=1)	Percutaneous transluminal coronary angioplasty

CHD=coronary heart disease; CVD=cardiovascular disease; MI=myocardial infarction; TIA=transient ischemic attack; AF=atrial fibrillation; ECG=electrocardiography; ICD=International Classification of Disease; PCI= percutaneous coronary intervention; CABG=Coronary artery bypass grafting; PTCA=Percutaneous transluminal coronary angioplasty.

Supplemental Table 5: Modelling method used to develop the prediction models.

Method	N (%)
Cox proportional hazards regression	160 (44%)
Accelerated failure time analysis	77 (21%)
Logistic regression	71 (20%)
Other parametric survival model	7 (2%)
Competing risk model	4 (1%)
Conditional logistic regression	2 (1%)
Poisson regression	2 (1%)
Expert weighing	1 (<0.5%)
Neural network	1 (<0.5%)
Other e.g. counted number of risk factors	2 (1%)
Not reported	36 (10%)
Total	363

Supplemental Table 6: Prediction horizons used for developed models.

Prediction horizon	N (%)
<5 years	3 (1%)
5 years	47 (13%)
5-10 years	25 (7%)
10 years	209 (58%)
10-20 years	14 (4%)
20-30 years	14 (4%)
>30 years	2 (1%)
Not reported	49 (13%)
Total	363

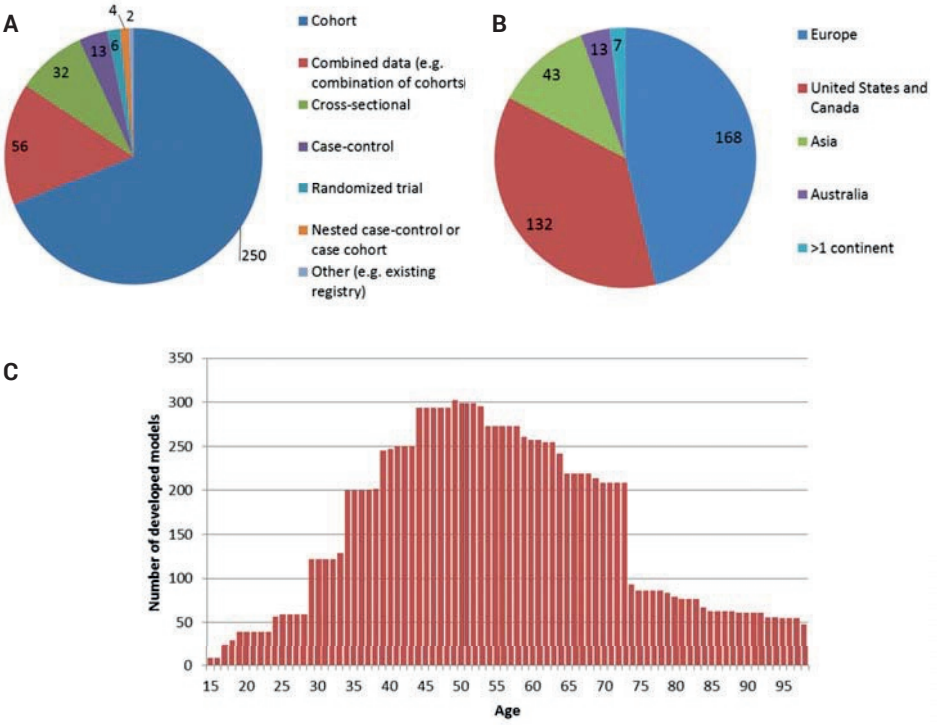
Supplemental Table 7: Characteristics of developed models that were and were not externally validated.

		Validated (n=132) N (%)	Not validated (n=231) N (%)
Study design	Longitudinal data (e.g. cohort)	123 (93%)	187 (81%)
	Cross-sectional data (e.g. case-control)	9 (7%)	44 (19%)
Gender	Men	52 (39%)	90 (39%)
	Women	41 (31%)	67 (29%)
	Men and women	39 (30%)	74 (32%)
Prediction horizon	<10 years	26 (20%)	49 (21%)
	10 years	88 (67%)	121 (52%)
	>10 years	4 (3%)	26 (11%)
	Not reported	14 (11%)	35 (15%)
Modelling method	Survival model	88 (67%)	149 (65%)
	Logistic regression	32 (24%)	39 (17%)
	Other	8 (6%)	11 (5%)
	Not reported	4 (3%)	22 (10%)
Internal validation	Yes	45 (34%)	35 (15%)
	No	87 (66%)	196 (85%)
Presentation	Model can be used for individual risk predictions	110 (83%)	161 (70%)
	Model cannot be used for individual risk predictions	22 (17%)	70 (30%)

Performance reported	Discrimination	46 (35%)	117 (51%)
	Calibration	34 (26%)	82 (35%)
	Overall performance	27 (20%)	8 (3%)
	Any performance measure	61 (46%)	130 (56%)

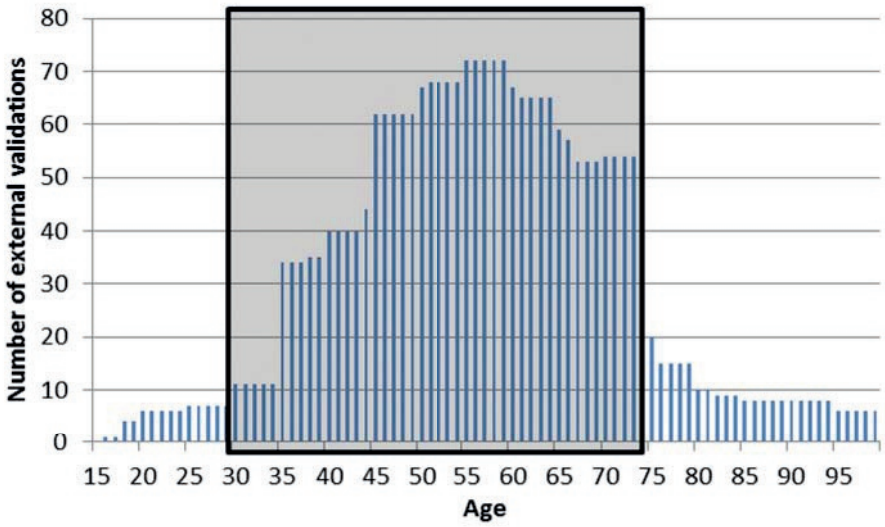
	N reported	Median	N reported	Median
Publication year	132	2003	231	2006
Impact factor	125	6.2	220	4.2
Number of participants	113	4,890	226	3,513
Number of events	80	364	209	181
Lower age limit	124	35	213	35
Upper age limit	124	74	213	74
Number of predictors	130	7	227	6

Supplemental figures

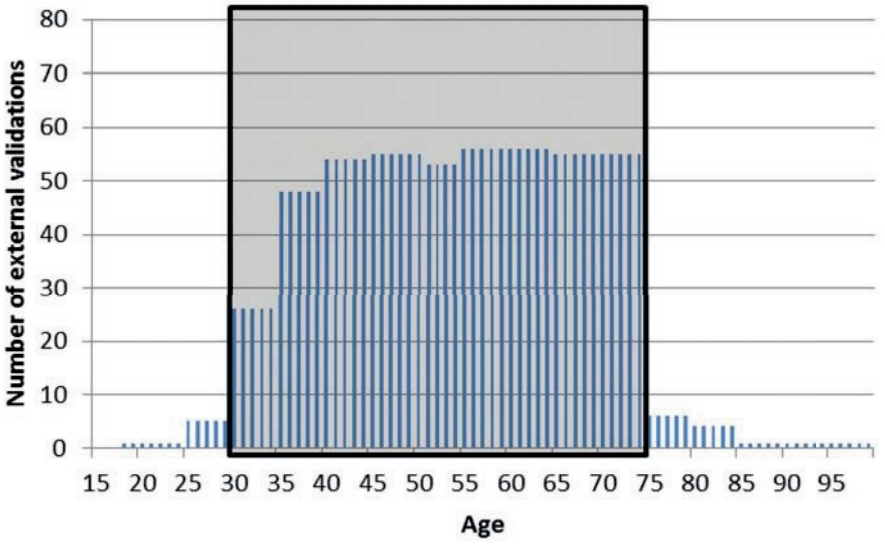


Supplemental Figure 1: Study design, location and age of included participants of all developed models. A: study design, B: location, C: age; bars indicate the number of models developed for that specific age. Models developed for e.g. age >16 were assumed to include people up to 99 years of age.

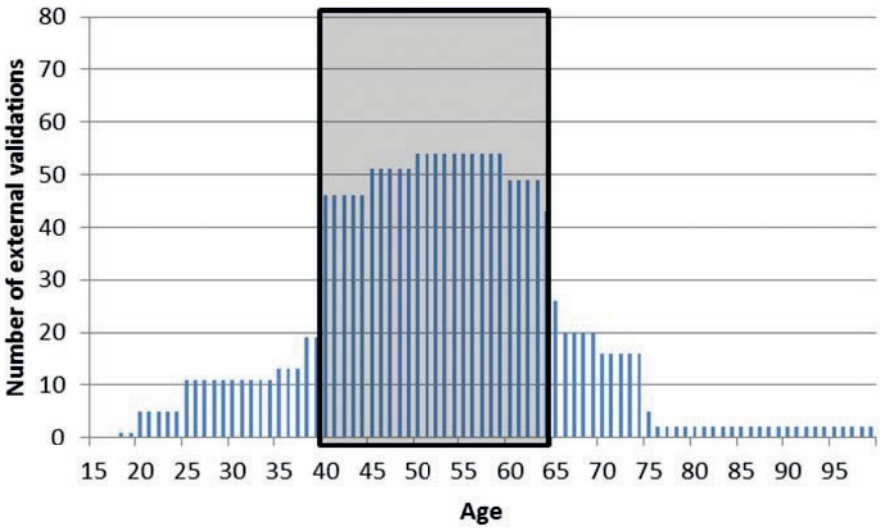
A


Number of external validations
0
10
20
30
40
50
60
70
80
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
Age

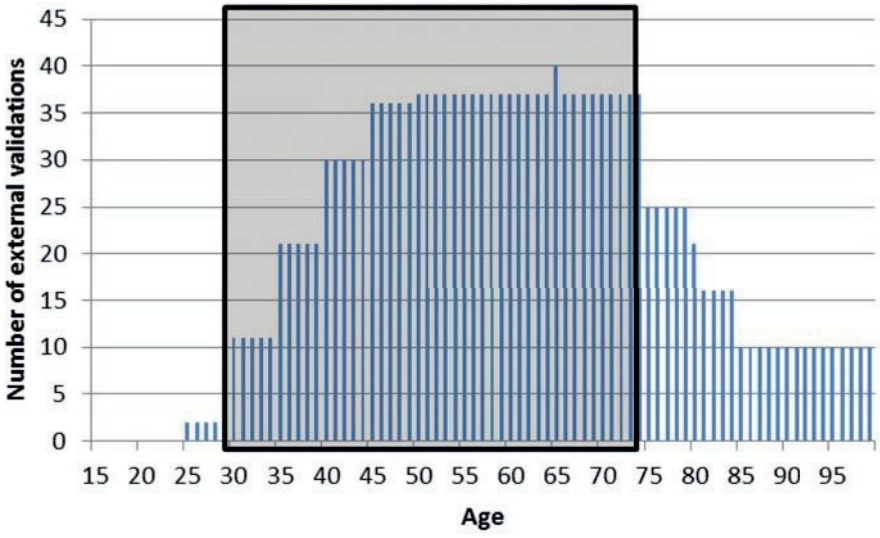
B


Number of external validations
0
10
20
30
40
50
60
70
80
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
Age

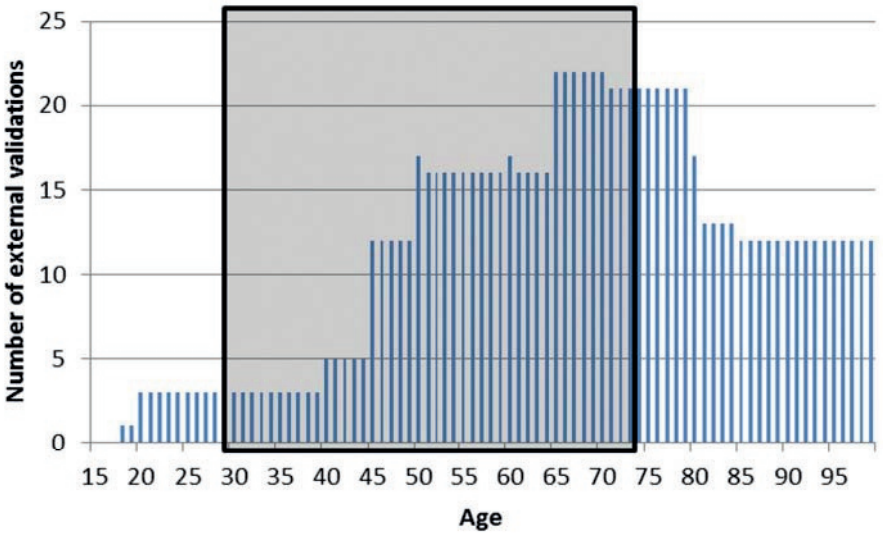
C


Number of external validations
0
10
20
30
40
50
60
70
80
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
Age

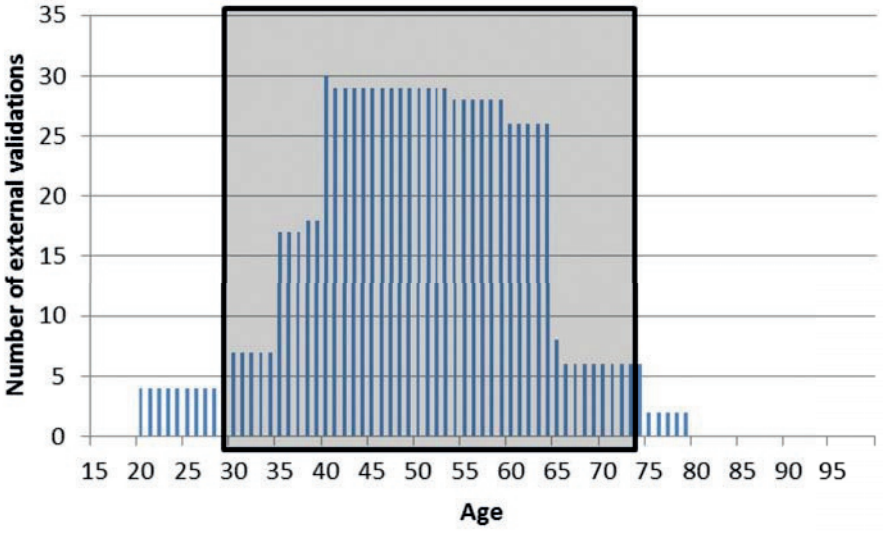
D


Number of external validations
0
5
10
15
20
25
30
35
40
45
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
Age

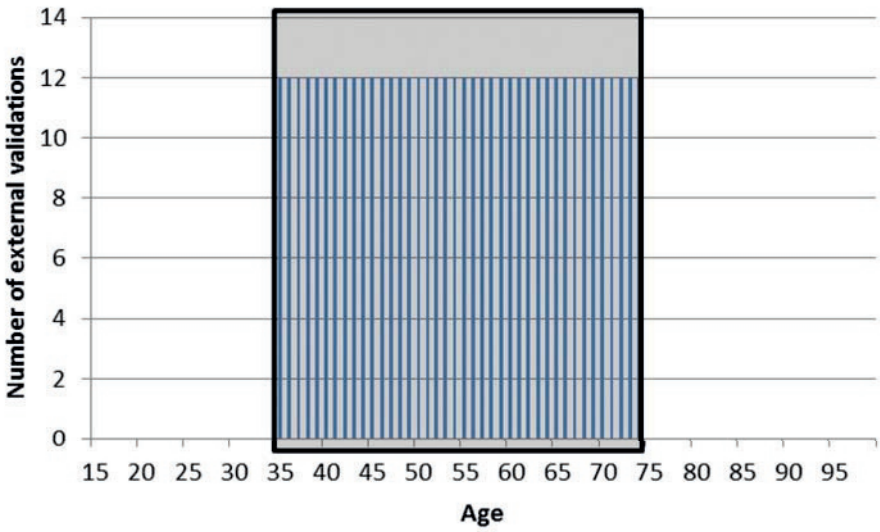
E


Number of external validations
0
5
10
15
20
25
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
Age

F


Number of external validations
0
5
10
15
20
25
30
35
15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
Age

G



Supplemental Figure 2: Ages of people included in external validations of the 7 most often validated models (see Table 3). The grey area indicates the age range in the original development study. A: Framingham Wilson 1998,[184] B: Framingham Anderson 1991a,[22] C: SCORE Conroy 2003,[87] D: Framingham D'Agostino 2008,[107] E: Framingham ATP III 2002,[1] F: Framingham Anderson 1991b,[51] G: QRISK Hippisley-Cox 2007.[34]

Reference list of included studies

1. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* 2002;106(25):3143-421.
2. Berry JD, Lloyd-Jones DM, Garside DB, Greenland P. Framingham risk score and prediction of coronary heart disease death in young men. *American Heart Journal* 2007;154(1):80-6.
3. Chamberlain AM, Agarwal SK, Folsom AR, Soliman EZ, Chambless LE, Crow R, et al. A clinical risk score for atrial fibrillation in a biracial prospective cohort (from the Atherosclerosis Risk in Communities [ARIC] study). *American Journal of Cardiology* 2011;107(1):85-91.
4. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women.[Summary for patients in Ann Intern Med. 2006 Jul 4;145(1):I19; PMID: 16818922]. *Annals of Internal Medicine* 2006;145(1):21-9.
5. Cooper JA, Miller GJ, Humphries SE. A comparison of the PROCAM and Framingham point-scoring systems for estimation of individual risk of coronary heart disease in the Second Northwick Park Heart Study. *Atherosclerosis* 2005;181(1):93-100.
6. Dhamoon MS, Moon YP, Paik MC, Sacco RL, Elkind MSV. The inclusion of stroke in risk stratification for primary prevention of vascular events: the Northern Manhattan Study. *Stroke* 2011;42(10):2878-82.
7. Dunder K, Lind L, Zethelius B, Berglund L, Lithell H. Evaluation of a scoring scheme, including proinsulin and the apolipoprotein B/apolipoprotein A1 ratio, for the risk of acute coronary events in middle-aged men: Uppsala Longitudinal Study of Adult Men (ULSAM). *American Heart Journal* 2004;148(4):596-601.
8. Duprez DA, Florea N, Zhong W, Grandits GA, Hawthorne CK, Hoke L, et al. Vascular and cardiac functional and structural screening to identify risk of future morbid events: preliminary observations. *J Am Soc Hypertens* 2011;5(5):401-9.
9. Kang HM, Kim D-J. Metabolic Syndrome versus Framingham Risk Score for Association of Self-Reported Coronary Heart Disease: The 2005 Korean Health and Nutrition Examination Survey. *Diabetes & Metabolism Journal* 2012;36(3):237-44.
10. Koller MT, Leening MJG, Wolbers M, Steyerberg EW, Hunink MGM, Schoop R, et al. Development and validation of a coronary risk prediction model for older U.S. and European persons in the cardiovascular health study and the Rotterdam Study. *Annals of Internal Medicine* 2012;157(6):389-97.
11. Koller MT, Steyerberg EW, Wolbers M, Stijnen T, Bucher HC, Hunink MGM, et al. Validity of the Framingham point scores in the elderly: results from the Rotterdam study. *American Heart Journal* 2007;154(1):87-93.

12. Murphy TP, Dhangana R, Pencina MJ, Zafar AM, D'Agostino RB. Performance of current guidelines for coronary heart disease prevention: optimal use of the Framingham-based risk assessment. *Atherosclerosis* 2011;216(2):452-7.
13. Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, Ridker PM. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. *Annals of Internal Medicine* 2009;150(2):65-72.
14. Paynter NP, Mazer NA, Pradhan AD, Gaziano JM, Ridker PM, Cook NR. Cardiovascular risk prediction in diabetic men and women using hemoglobin A1c vs diabetes as a high-risk equivalent. *Archives of Internal Medicine* 2011;171(19):1712-8.
15. Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score.[Erratum appears in JAMA. 2007 Apr 4;297(13):1433]. *JAMA* 2007;297(6):611-9.
16. Ridker PM, Paynter NP, Rifai N, Gaziano JM, Cook NR. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. *Circulation* 2008;118(22):2243-51, 4p following 51.
17. Rifkin DE, Ix JH, Wassel CL, Criqui MH, Allison MA. Renal artery calcification and mortality among clinically asymptomatic adults. *Journal of the American College of Cardiology* 2012;60(12):1079-85.
18. Simons LA, Simons J, Friedlander Y, McCallum J, Palaniappan L. Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study. *Med J Aust* 2003;178(3):113-6.
19. Veeranna V, Zalawadiya SK, Niraj A, Pradhan J, Ference B, Burack RC, et al. Homocysteine and reclassification of cardiovascular disease risk. *Journal of the American College of Cardiology* 2011;58(10):1025-33.
20. Cook NR, Paynter NP, Eaton CB, Manson JE, Martin LW, Robinson JG, et al. Comparison of the Framingham and Reynolds Risk scores for global cardiovascular risk prediction in the multiethnic Women's Health Initiative. *Circulation* 2012;125(14):1748-56, S1-11.
21. Alssema M, Newson RS, Bakker SJL, Stehouwer CDA, Heymans MW, Nijpels G, et al. One risk assessment tool for cardiovascular disease, type 2 diabetes, and chronic kidney disease. *Diabetes Care* 2012;35(4):741-8.
22. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. *Am Heart J* 1991;121(1 Pt 2):293-8.
23. Bastuji-Garin S, Deverly A, Moyse D, Castaigne A, Mancia G, de Leeuw PW, et al. The Framingham prediction rule is not valid in a European population of treated hypertensive patients. *Journal of hypertension* 2002;20(10):1973-80.

24. Bhopal R, Fischbacher C, Vartiainen E, Unwin N, White M, Alberti G. Predicted and observed cardiovascular disease in South Asians: application of FINRISK, Framingham and SCORE models to Newcastle Heart Project data. *Journal of Public Health* 2005;27(1):93-100.
25. Brindle P, Emberson J, Lampe F, Walker M, Whincup P, Fahey T, et al. Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study. *BMJ* 2003;327(7426):1267.
26. Brindle PM, McConnachie A, Upton MN, Hart CL, Davey Smith G, Watt GCM. The accuracy of the Framingham risk-score in different socioeconomic groups: a prospective study. *British Journal of General Practice* 2005;55(520):838-45.
27. Chen L, Tonkin AM, Moon L, Mitchell P, Dobson A, Giles G, et al. Recalibration and validation of the SCORE risk chart in the Australian population: the AusSCORE chart. *European Journal of Cardiovascular Prevention & Rehabilitation* 2009;16(5):562-70.
28. Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ* 2009;339:b2584.
29. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ (Clinical research ed.)* 2010;340:c2442.
30. Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. *BMJ* 2012;344:e4181.
31. de Ruijter W, Westendorp RGJ, Assendelft WJJ, den Elzen WPJ, de Craen AJM, le Cessie S, et al. Use of Framingham risk score and new biomarkers to predict cardiovascular mortality in older people: population based observational cohort study. *BMJ* 2009;338:a3083.
32. Hense HW, Schulte H, Lowel H, Assmann G, Keil U. Framingham risk function overestimates risk of coronary heart disease in men and women from Germany--results from the MONICA Augsburg and the PROCAM cohorts. *Eur Heart J* 2003;24(10):937-45.
33. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart* 2008;94(1):34-9.
34. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007;335(7611):136.
35. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336(7659):1475-82.

36. Marshall T. Identification of patients for clinical risk assessment by prediction of cardiovascular risk using default risk factor values. *BMC Public Health* 2008;8:25.
37. May M, Lawlor DA, Brindle P, Patel R, Ebrahim S. Cardiovascular disease risk assessment in older women: can we improve on Framingham? British Women's Heart and Health prospective cohort study. *Heart* 2006;92(10):1396-401.
38. Milne R, Gamble G, Whitlock G, Jackson R. Discriminative ability of a risk-prediction tool derived from the Framingham Heart Study compared with single risk factors. *The New Zealand medical journal* 2003;116(1185):U663.
39. Milne R, Gamble G, Whitlock G, Jackson R. Framingham Heart Study risk equation predicts first cardiovascular event rates in New Zealanders at the population level. *New Zealand Medical Journal* 2003;116(1185):U662.
40. Nelson MR, Ramsay E, Ryan P, Willson K, Tonkin AM, Wing L, et al. A score for the prediction of cardiovascular events in the hypertensive aged. *American Journal of Hypertension* 2012;25(2):190-4.
41. Nelson MR, Ryan P, Tonkin AM, Ramsay E, Willson K, Wing LWH, et al. Prediction of cardiovascular events in subjects in the second Australian National Blood Pressure study. *Hypertension* 2010;56(1):44-8.
42. Orford JL, Sesso HD, Stedman M, Gagnon D, Vokonas P, Gaziano JM. A comparison of the Framingham and European Society of Cardiology coronary heart disease risk prediction models in the normative aging study. *Am Heart J* 2002;144(1):95-100.
43. Pandya A, Weinstein MC, Gaziano TA. A comparative assessment of non-laboratory-based versus commonly used laboratory-based cardiovascular disease risk scores in the NHANES III population. *PLoS ONE [Electronic Resource]* 2011;6(5):e20416.
44. Ramsay SE, Morris RW, Whincup PH, Papacosta AO, Thomas MC, Wannamethee SG. Prediction of coronary heart disease risk by Framingham and SCORE risk assessments varies by socioeconomic position: results from a study in British men. *European Journal of Cardiovascular Prevention & Rehabilitation* 2011;18(2):186-93.
45. Riddell T, Wells S, Jackson R, Lee A-W, Crengle S, Bramley D, et al. Performance of Framingham cardiovascular risk scores by ethnic groups in New Zealand: PREDICT CVD-*New Zealand Medical Journal* 2010;123(1309):50-61.
46. Tunstall-Pedoe H, Woodward M, estimation Sgor. By neglecting deprivation, cardiovascular risk scoring will exacerbate social gradients in disease. *Heart* 2006;92(3):307-10.
47. Villines TC, Taylor AJ. Multi-ethnic study of atherosclerosis arterial age versus framingham 10-year or lifetime cardiovascular risk. *American Journal of Cardiology* 2012;110(11):1627-30.
48. Woodward M, Brindle P, Tunstall-Pedoe H, estimation Sgor. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart* 2007;93(2):172-6.

49. Zhang X-F, Attia J, D'Este C, Yu X-H, Wu X-G. A risk score predicted coronary heart disease and stroke in a Chinese cohort. *Journal of Clinical Epidemiology* 2005;58(9):951-8.
50. Zomer E, Owen A, Magliano DJ, Liew D, Reid C. Validation of two Framingham cardiovascular risk prediction algorithms in an Australian population: the 'old' versus the 'new' Framingham equation. *European Journal of Cardiovascular Prevention & Rehabilitation* 2011;18(1):115-20.
51. Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk profile. A statement for health professionals. *Circulation* 1991;83(1):356-62.
52. Boudik F, Reissigova J, Hrach K, Tomeckova M, Bultas J, Anger Z, et al. Primary prevention of coronary artery disease among middle aged men in Prague: twenty-year follow-up results. *Atherosclerosis* 2006;184(1):86-93.
53. de la Iglesia B, Potter JF, Poulter NR, Robins MM, Skinner J. Performance of the ASSIGN cardiovascular disease risk score on a UK cohort of patients from general practice. *Heart* 2011;97(6):491-9.
54. Jimenez-Corona A, Lopez-Ridaura R, Williams K, Gonzalez-Villalpando ME, Simon J, Gonzalez-Villalpando C. Applicability of Framingham risk equations for studying a low-income Mexican population. [Spanish]Aplicabilidad del puntaje de Framingham en poblacion mexicana de nivel socioeconomico bajo. Salud Publica de Mexico 2009;51(4):298-305.
55. Ketola E, Laatikainen T, Vartiainen E. Evaluating risk for cardiovascular diseases--vain or value? How do different cardiovascular risk scores act in real life. *European Journal of Public Health* 2010;20(1):107-12.
56. Laurier D, Nguyen PC, Cazelles B, Segond P. Estimation of CHD risk in a French working population using a modified Framingham model. The PCV-METRA Group. *J Clin Epidemiol* 1994;47(12):1353-64.
57. Ramachandran S, French JM, Vanderpump MP, Croft P, Neary RH. Using the Framingham model to predict heart disease in the United Kingdom: retrospective study. *BMJ* 2000;320(7236):676-7.
58. Wang Z, Hoy WE. Is the Framingham coronary heart disease absolute risk function applicable to Aboriginal people? *Med J Aust* 2005;182(2):66-9.
59. Wannamethee SG, Shaper AG, Lennon L, Morris RW. Metabolic syndrome vs Framingham Risk Score for prediction of coronary heart disease, stroke, and type 2 diabetes mellitus. *Archives of Internal Medicine* 2005;165(22):2644-50.
60. Arima H, Yonemoto K, Doi Y, Ninomiya T, Hata J, Tanizaki Y, et al. Development and validation of a cardiovascular risk prediction model for Japanese: the Hisayama study. *Hypertension Research - Clinical & Experimental* 2009;32(12):1119-22.
61. Asayama K, Ohkubo T, Sato A, Hara A, Obara T, Yasui D, et al. Proposal of a risk-stratification system for the Japanese population based on blood pressure levels: the Ohasama study. *Hypertension Research - Clinical & Experimental* 2008;31(7):1315-22.

62. Asia Pacific Cohort Studies Collaboration. Coronary risk prediction for those with and without diabetes. *Eur J Cardiovasc Prev Rehabil* 2006;13(1):30-6.
63. Asia Pacific Cohort Studies C, Barzi F, Patel A, Gu D, Sritara P, Lam TH, et al. Cardiovascular risk prediction tools for populations in Asia. *Journal of Epidemiology & Community Health* 2007;61(2):115-21.
64. Aslibekyan S, Campos H, Loucks EB, Linkletter CD, Ordovas JM, Baylin A. Development of a cardiovascular risk score for use in low- and middle-income countries. *Journal of Nutrition* 2011;141(7):1375-80.
65. Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. *Circulation* 2002;105(3):310-5.
66. Assmann G, Schulte H, Seedorf U. Cardiovascular risk assessment in the metabolic syndrome: results from the Prospective Cardiovascular Munster (PROCAM) Study. *International Journal of Obesity* 2008;32 Suppl 2:S11-6.
67. Chien KL, Hsu HC, Su TC, Chang WT, Chen PC, Sung FC, et al. Constructing a point-based prediction model for the risk of coronary artery disease in a Chinese community: A report from a cohort study in Taiwan. *International Journal of Cardiology* 2012;157(2):263-68.
68. Empana JP, Ducimetiere P, Arveiler D, Ferrieres J, Evans A, Ruidavets JB, et al. Are the Framingham and PROCAM coronary heart disease risk functions applicable to different European populations? The PRIME Study. *Eur Heart J* 2003;24(21):1903-11.
69. Ferrario M, Chiodini P, Chambless LE, Cesana G, Vanuzzo D, Panico S, et al. Prediction of coronary events in a low incidence population. Assessing accuracy of the CUORE Cohort Study prediction equation. *Int J Epidemiol* 2005;34(2):413-21.
70. Assmann G, Schulte H, Cullen P, Seedorf U. Assessing risk of myocardial infarction and stroke: new data from the Prospective Cardiovascular Munster (PROCAM) study. *European Journal of Clinical Investigation* 2007;37(12):925-32.
71. Chien KL, Su TC, Hsu HC, Chang WT, Chen PC, Sung FC, et al. Constructing the prediction model for the risk of stroke in a Chinese population: report from a cohort study in Taiwan. *Stroke* 2010;41(9):1858-64.
72. Balkau B, Hu G, Qiao Q, Tuomilehto J, Borch-Johnsen K, Pyorala K, et al. Prediction of the risk of cardiovascular mortality using a score that includes glucose as a risk factor. The DECODE Study. *Diabetologia* 2004;47(12):2118-28.
73. Beer C, Alfonso H, Flicker L, Norman PE, Hankey GJ, Almeida OP. Traditional risk factors for incident cardiovascular events have limited importance in later life compared with the health in men study cardiovascular risk score. *Stroke* 2011;42(4):952-9.
74. Bell K, Hayen A, McGeechan K, Neal B, Irwig L. Effects of additional blood pressure and lipid measurements on the prediction of cardiovascular risk. *European Journal of Preventive Cardiology* 2012;19(6):1474-85.

75. Berard E, Bongard V, Arveiler D, Amouyel P, Wagner A, Dallongeville J, et al. Ten-year risk of all-cause mortality: assessment of a risk prediction algorithm in a French general population. *European Journal of Epidemiology* 2011;26(5):359-68.
76. Boland B, De Muylder R, Goderis G, Degryse J, Gueuning Y, Paulus D, et al. Cardiovascular prevention in general practice: development and validation of an algorithm. *Acta Cardiologica* 2004;59(6):598-605.
77. Bolton JL, Stewart MCW, Wilson JF, Anderson N, Price JF. Improvement in Prediction of Coronary Heart Disease Risk over Conventional Risk Factors Using SNPs Identified in Genome-Wide Association Studies. *PLoS ONE* 2013;8(2).
78. Brand RJ, Rosenman RH, Sholtz RI, Friedman M. Multivariate prediction of coronary heart disease in the Western Collaborative Group Study compared to the findings of the Framingham study. *Circulation* 1976;53(2):348-55.
79. Braun J, Bopp M, Faeh D. Blood glucose may be an alternative to cholesterol in CVD risk prediction charts. *Cardiovascular Diabetology* 2013;12(1).
80. Brautbar A, Ballantyne CM, Lawson K, Nambi V, Chambless L, Folsom AR, et al. Impact of adding a single allele in the 9p21 locus to traditional risk factors on reclassification of coronary heart disease risk and implications for lipid-modifying therapy in the Atherosclerosis Risk in Communities study. *Circulation. Cardiovascular Genetics* 2009;2(3):279-85.
81. Brindle P, May M, Gill P, Cappuccio F, D'Agostino R, Sr., Fischbacher C, et al. Primary prevention of cardiovascular disease: a web-based risk score for seven British black and minority ethnic groups. *Heart* 2006;92(11):1595-602.
82. Chambless LE, Folsom AR, Sharrett AR, Sorlie P, Couper D, Szklo M, et al. Coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC) study. *J Clin Epidemiol* 2003;56(9):880-90.
83. Nambi V, Boerwinkle E, Lawson K, Brautbar A, Chambless L, Franeschini N, et al. The 9p21 genetic variant is additive to carotid intima media thickness and plaque in improving coronary heart disease risk prediction in white participants of the Atherosclerosis Risk in Communities (ARIC) Study. *Atherosclerosis* 2012;222(1):135-7.
84. Nambi V, Chambless L, Folsom AR, He M, Hu Y, Mosley T, et al. Carotid intima-media thickness and presence or absence of plaque improves prediction of coronary heart disease risk: the ARIC (Atherosclerosis Risk In Communities) study. *Journal of the American College of Cardiology* 2010;55(15):1600-7.
85. Nambi V, Chambless L, He M, Folsom AR, Mosley T, Boerwinkle E, et al. Common carotid artery intima-media thickness is as good as carotid intima-media thickness of all carotid artery segments in improving prediction of coronary heart disease risk in the Atherosclerosis Risk in Communities (ARIC) study. *European Heart Journal* 2012;33(2):183-90.

86. Ciampi A, Courteau J, Niyonsenga T, Xhignesse M, Lussier-Cacan S, Roy M. Family history and the risk of coronary heart disease: comparing predictive models. *Eur J Epidemiol* 2001;17(7):609-20.
87. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J* 2003;24(11):987-1003.
88. Aktas MK, Ozduran V, Pothier CE, Lang R, Lauer MS. Global risk scores and exercise testing for predicting all-cause mortality in a preventive medicine program. *JAMA* 2004;292(12):1462-8.
89. Barroso LC, Muro EC, Herrera ND, Ochoa GF, Hueros JIC, Buitrago F. Performance of the Framingham and SCORE cardiovascular risk prediction functions in a non-diabetic population of a Spanish health care centre: a validation study. *Scandinavian Journal of Primary Health Care* 2010;28(4):242-8.
90. Comin E, Solanas P, Cabezas C, Subirana I, Ramos R, Gene-Badia J, et al. Estimating cardiovascular risk in Spain using different algorithms. [Spanish]Rendimiento de la estimacion del riesgo cardiovascular en Espana mediante la utilizacion de distintas funciones. *Revista Espanola de Cardiologia* 2007;60(7):693-702.
91. De Bacquer D, De Backer G. Predictive ability of the SCORE Belgium risk chart for cardiovascular mortality. *International Journal of Cardiology* 2010;143(3):385-90.
92. Lindman AS, Veierod MB, Pedersen JI, Tverdal A, Njolstad I, Selmer R. The ability of the SCORE high-risk model to predict 10-year cardiovascular disease mortality in Norway. *European Journal of Cardiovascular Prevention & Rehabilitation* 2007;14(4):501-7.
93. Mainous AG, 3rd, Koopman RJ, Diaz VA, Everett CJ, Wilson PWF, Tilley BC. A coronary heart disease risk score based on patient-reported information. *American Journal of Cardiology* 2007;99(9):1236-41.
94. Merry AHH, Boer JMA, Schouten LJ, Ambergen T, Steyerberg EW, Feskens EJM, et al. Risk prediction of incident coronary heart disease in The Netherlands: re-estimation and improvement of the SCORE risk function. *European Journal of Preventive Cardiology* 2012;19(4):840-8.
95. Nielsen M, Ganz M, Lauze F, Pettersen PC, de Bruijne M, Clarkson TB, et al. Distribution, size, shape, growth potential and extent of abdominal aortic calcified deposits predict mortality in postmenopausal women. *BMC Cardiovascular Disorders* 2010;10:56.
96. Panagiotakos DB, Fitzgerald AP, Pitsavos C, Pipilis A, Graham I, Stefanadis C. Statistical modelling of 10-year fatal cardiovascular disease risk in Greece: the HellenicSCORE (a calibration of the ESC SCORE project). *Hjc Hellenic Journal of Cardiology* 2007;48(2):55-63.
97. Saidj M, Jorgensen T, Prescott E, Borglykke A. Poor predictive ability of the risk chart SCORE in a Danish population. *Danish Medical Journal* 2013;60(5).

98. Scheltens T, Verschuren WMM, Boshuizen HC, Hoes AW, Zuithoff NP, Bots ML, et al. Estimation of cardiovascular risk: a comparison between the Framingham and the SCORE model in people under 60 years of age. *European Journal of Cardiovascular Prevention & Rehabilitation* 2008;15(5):562-6.
99. Stenlund H, Lonnberg G, Jenkins P, Norberg M, Persson M, Messner T, et al. Fewer deaths from cardiovascular disease than expected from the Systematic Coronary Risk Evaluation chart in a Swedish population. *European Journal of Cardiovascular Prevention & Rehabilitation* 2009;16(3):321-4.
100. Ulmer H, Kollerits B, Kelleher C, Diem G, Concin H. Predictive accuracy of the SCORE risk function for cardiovascular disease in clinical practice: a prospective evaluation of 44 649 Austrian men and women. *European Journal of Cardiovascular Prevention & Rehabilitation* 2005;12(5):433-41.
101. van der Heijden AAWA, Ortegon MM, Niessen LW, Nijpels G, Dekker JM. Prediction of coronary heart disease risk in a general, pre-diabetic, and diabetic population during 10 years of follow-up: accuracy of the Framingham, SCORE, and UKPDS risk functions: The Hoorn Study. *Diabetes Care* 2009;32(11):2094-8.
102. 1van Dis I, Kromhout D, Geleijnse JM, Boer JMA, Verschuren WMM. Evaluation of cardiovascular risk predicted by different SCORE equations: the Netherlands as an example. *European Journal of Cardiovascular Prevention & Rehabilitation* 2010;17(2):244-9.
103. Cross DS, McCarty CA, Hytopoulos E, Beggs M, Nolan N, Harrington DS, et al. Coronary risk assessment among intermediate risk patients using a clinical and biomarker based algorithm developed and validated in two population cohorts. *Current Medical Research & Opinion* 2012;28(11):1819-30.
104. D'Agostino RB, Wolf PA, Belanger AJ, Kannel WB. Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study. *Stroke* 1994;25(1):40-3.
105. Bineau S, Dufouil C, Helmer C, Ritchie K, Empana J-P, Ducimetiere P, et al. Framingham stroke risk function in a large population-based cohort of elderly people: the 3C study. *Stroke* 2009;40(5):1564-70.
106. D'Agostino RB, Russell MW, Huse DM, Ellison RC, Silbershatz H, Wilson PW, et al. Primary and subsequent coronary risk appraisal: new results from the Framingham study. *Am Heart J* 2000;139(2 Pt 1):272-81.
107. D'Agostino RB, Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008;117(6):743-53.
108. Bozorgmanesh M, Hadaegh F, Azizi F. Predictive accuracy of the 'Framingham's general CVD algorithm' in a Middle Eastern population: Tehran Lipid and Glucose Study. *International Journal of Clinical Practice* 2011;65(3):264-73.

109. Chamnan P, Simmons RK, Hori H, Sharp S, Khaw K-T, Wareham NJ, et al. A simple risk score using routine data for predicting cardiovascular disease in primary care. *British Journal of General Practice* 2010;60(577):e327-34.
110. Hamer M, Chida Y, Stamatakis E. Utility of C-reactive protein for cardiovascular risk stratification across three age groups in subjects without existing cardiovascular diseases. *American Journal of Cardiology* 2009;104(4):538-42.
111. Hurley LP, Dickinson LM, Estacio RO, Steiner JF, Havranek EP. Prediction of cardiovascular death in racial/ethnic minorities using Framingham risk factors. *Circulation. Cardiovascular Quality & Outcomes* 2010;3(2):181-7.
112. Khalili D, Hadaegh F, Soori H, Steyerberg EW, Bozorgmanesh M, Azizi F. Clinical usefulness of the Framingham cardiovascular risk profile beyond its statistical performance: the Tehran Lipid and Glucose Study. *American Journal of Epidemiology* 2012;176(3):177-86.
113. Mohammadreza B, Farzad H, Davoud K, Fereidoun Prof AF. Prognostic significance of the complex "Visceral Adiposity Index" vs. simple anthropometric measures: Tehran lipid and glucose study. *Cardiovascular Diabetology* 2012;11:20.
114. Simmons RK, Coleman RL, Price HC, Holman RR, Khaw K-T, Wareham NJ, et al. Performance of the UK Prospective Diabetes Study Risk Engine and the Framingham Risk Equations in Estimating Cardiovascular Disease in the EPIC-Norfolk Cohort. *Diabetes Care* 2009;32(4):708-13.
115. Ito H, Pacold IV, Durazo-Arvizu R, Liu K, Shilipak MG, Goff DC, Jr., et al. The effect of including cystatin C or creatinine in a cardiovascular risk model for asymptomatic individuals: the multi-ethnic study of atherosclerosis. *American Journal of Epidemiology* 2011;174(8):949-57.
116. Davies RW, Dandona S, Stewart AFR, Chen L, Ellis SG, Tang WHW, et al. Improved prediction of cardiovascular disease based on a panel of single nucleotide polymorphisms identified through genome-wide association studies. *Circulation. Cardiovascular Genetics* 2010;3(5):468-74.
117. Donfrancesco C, Palmieri L, Cooney M-T, Vanuzzo D, Panico S, Cesana G, et al. Italian cardiovascular mortality charts of the CUORE project: are they comparable with the SCORE charts? *European Journal of Cardiovascular Prevention & Rehabilitation* 2010;17(4):403-9.
118. Empana JP, Tafflet M, Escolano S, Vergnaux AC, Bineau S, Ruidavets JB, et al. Predicting CHD risk in France: A pooled analysis of the D.E.S.I.R., Three City, PRIME, and SU.VI.MAX studies. *European Journal of Cardiovascular Prevention and Rehabilitation* 2011;18(2):175-85.
119. Faeh D, Braun J, Rufibach K, Puhan MA, Marques-Vidal P, Bopp M. Population Specific and Up to Date Cardiovascular Risk Charts Can Be Efficiently Obtained with Record Linkage of Routine and Observational Data. *PLoS ONE* 2013;8(2).

120. Folsom AR, Chambless LE, Duncan BB, Gilbert AC, Pankow JS. Prediction of coronary heart disease in middle-aged adults with diabetes. *Diabetes Care* 2003;26(10):2777-84.
121. Friedland DR, Cederberg C, Tarima S. Audiometric pattern as a predictor of cardiovascular status: development of a model for assessment of risk. *Laryngoscope* 2009;119(3):473-86.
122. Gaziano TA, Young CR, Fitzmaurice G, Atwood S, Gaziano JM. Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the NHANES I Follow-up Study cohort. *Lancet* 2008;371(9616):923-31.
123. Glynn RJ, L'Italien GJ, Sesso HD, Jackson EA, Buring JE. Development of predictive models for long-term cardiovascular risk associated with systolic and diastolic blood pressure. *Hypertension* 2002;39(1):105-10.
124. Hesse B, Morise A, Pothier CE, Blackstone EH, Lauer MS. Can we reliably predict long-term mortality after exercise testing? An external validation. *American Heart Journal* 2005;150(2):307-14.
125. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. *BMJ* 2010;341:c6624.
126. Hoes AW, Grobbee DE, Valkenburg HA, Lubsen J, Hofman A. Cardiovascular risk and all-cause mortality; a 12 year follow-up study in The Netherlands. *Eur J Epidemiol* 1993;9(3):285-92.
127. Houterman S, Boshuizen HC, Verschuren WM, Giampaoli S, Nissinen A, Menotti A, et al. Predicting cardiovascular risk in the elderly in different European countries. *Eur Heart J* 2002;23(4):294-300.
128. Ishikawa S, Matsumoto M, Kayaba K, Gotoh T, Nago N, Tsutsumi A, et al. Risk charts illustrating the 10-year risk of stroke among residents of Japanese rural communities: the JMS Cohort Study. *Journal of Epidemiology* 2009;19(2):101-6.
129. Janssen I, Katzmarzyk PT, Church TS, Blair SN. The Cooper Clinic Mortality Risk Index: clinical score sheet for men. *American Journal of Preventive Medicine* 2005;29(3):194-203.
130. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. *Am J Cardiol* 1976;38(1):46-51.
131. Keys A, Aravanis C, Blackburn H, Van Buchem FS, Buzina R, Djordjevic BS, et al. Probability of middle-aged men developing coronary heart disease in five years. *Circulation* 1972;45(4):815-28.
132. Knuiman MW, Vu HT. Prediction of coronary heart disease mortality in Busselton, Western Australia: an evaluation of the Framingham, national health epidemiologic follow up study, and WHO ERICA risk scores. *J Epidemiol Community Health* 1997;51(5):515-9.

133. Knuiman MW, Vu HT, Bartholomew HC. Multivariate risk estimation for coronary heart disease: the Busselton Health Study. *Aust N Z J Public Health* 1998;22(7):747-53.
134. L'Italien G, Ford I, Norrie J, LaPuerta P, Ehreth J, Jackson J, et al. The cardiovascular event reduction tool (CERT)--a simplified cardiac risk prediction model developed from the West of Scotland Coronary Prevention Study (WOSCOPS). *Am J Cardiol* 2000;85(6):720-4.
135. Larson MG. Assessment of cardiovascular risk factors in the elderly: the Framingham Heart Study. *Stat Med* 1995;14(16):1745-56.
136. Leaverton PE, Sorlie PD, Kleinman JC, Dannenberg AL, Ingster-Moore L, Kannel WB, et al. Representativeness of the Framingham risk model for coronary heart disease mortality: a comparison with a national cohort study. *J Chronic Dis* 1987;40(8):775-84.
137. Lee ET, Howard BV, Wang W, Welty TK, Galloway JM, Best LG, et al. Prediction of coronary heart disease in a population with high prevalence of diabetes and albuminuria: the Strong Heart Study. *Circulation* 2006;113(25):2897-905.
138. Lee J, Heng D, Ma S, Chew S-K, Hughes K, Tai ES. The metabolic syndrome and mortality: the Singapore Cardiovascular Cohort Study. *Clinical Endocrinology* 2008;69(2):225-30.
139. Levy D, Wilson PW, Anderson KM, Castelli WP. Stratifying the patient at risk from coronary disease: new insights from the Framingham Heart Study. *Am Heart J* 1990;119(3 Pt 2):712-7; discussion 17.
140. Liu J, Hong Y, D'Agostino RB, Sr., Wu Z, Wang W, Sun J, et al. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study. *JAMA* 2004;291(21):2591-9.
141. Lloyd-Jones DM, Leip EP, Larson MG, D'Agostino RB, Beiser A, Wilson PW, et al. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. *Circulation* 2006;113(6):791-8.
142. Lumley T, Kronmal RA, Cushman M, Manolio TA, Goldstein S. A stroke prediction score in the elderly: validation and Web-based application. *J Clin Epidemiol* 2002;55(2):129-36.
143. Voko Z, Hollander M, Koudstaal PJ, Hofman A, Breteler MMB. How do American stroke risk functions perform in a Western European population? *Neuroepidemiology* 2004;23(5):247-53.
144. Macfarlane PW, Norrie J. The value of the electrocardiogram in risk assessment in primary prevention: Experience from the West of Scotland Coronary Prevention Study. *Journal of Electrocardiology* 2007;40(1):101-09.
145. Mannan H, Stevenson C, Peeters A, Walls H, McNeil J. Framingham risk prediction equations for incidence of cardiovascular disease using detailed measures for smoking. *Heart International* 2010;5(2):e11.

146. Mannan HR, Stevenson CE, Peeters A, Walls HL, McNeil JJ. Age at quitting smoking as a predictor of risk of cardiovascular disease incidence independent of smoking status, time since quitting and pack-years. *BMC Research Notes* 2011;4:39.
147. Mannan HR, Stevenson CE, Peeters A, McNeil JJ. A new set of risk equations for predicting long term risk of all-cause mortality using cardiovascular risk factors. *Preventive Medicine* 2013;56(1):41-45.
148. Matsumoto M, Ishikawa S, Kayaba K, Gotoh T, Nago N, Tsutsumi A, et al. Risk charts illustrating the 10-year risk of myocardial infarction among residents of Japanese rural communities: the JMS Cohort Study. *Journal of Epidemiology* 2009;19(2):94-100.
149. May M, Sterne JAC, Shipley M, Brunner E, d'Agostino R, Whincup P, et al. A coronary heart disease risk model for predicting the effect of potent antiretroviral therapy in HIV-1 infected men. *International Journal of Epidemiology* 2007;36(6):1309-18.
150. McGorrian C, Yusuf S, Islam S, Jung H, Rangarajan S, Avezum A, et al. Estimating modifiable coronary heart disease risk in multiple regions of the world: the INTERHEART Modifiable Risk Score. *European Heart Journal* 2011;32(5):581-9.
151. McNeil JJ, Peeters A, Liew D, Lim S, Vos T. A model for predicting the future incidence of coronary heart disease within percentiles of coronary heart disease risk. *J Cardiovasc Risk* 2001;8(1):31-7.
152. Menotti A, Keys A, Kromhout D, Nissinen A, Blackburn H, Fidanza F, et al. Twenty-five-year mortality from coronary heart disease and its prediction in five cohorts of middle-aged men in Finland, The Netherlands, and Italy. *Prev Med* 1990;19(3):270-8.
153. Menotti A, Farchi G, Seccareccia F. The prediction of coronary heart disease mortality as a function of major risk factors in over 30 000 men in the Italian RIFLE pooling Project. A comparison with the MRFIT primary screenees. The RIFLE research group. *J Cardiovasc Risk* 1994;1(3):263-70.
154. Menotti A, Lanti M, Puddu PE, Mancini M, Zanchetti A, Cirillo M, et al. First risk functions for prediction of coronary and cardiovascular disease incidence in the Gubbio Population Study. *Ital Heart J* 2000;1(6):394-9.
155. Menotti A, Lanti M, Puddu PE, Carratelli L, Mancini M, Motolese M, et al. The risk functions incorporated in Riscard 2002: a software for the prediction of cardiovascular risk in the general population based on Italian data. *Ital Heart J* 2002;3(2):114-21.
156. Menotti A, Lanti M, Agabiti-Rosei E, Carratelli L, Cavera G, Dormi A, et al. Riskard 2New tools for prediction of cardiovascular disease risk derived from Italian population studies. *Nutrition Metabolism & Cardiovascular Diseases* 2005;15(6):426-40.
157. Moons KG, Bots ML, Salonen JT, Elwood PC, Freire de Concalves A, Nikitin Y, et al. Prediction of stroke in the general population in Europe (EUROSTROKE): Is there a role for fibrinogen and electrocardiography? *J Epidemiol Community Health* 2002;56 Suppl 1:i30-6.

158. Group NDR. Risk assessment chart for death from cardiovascular disease based on a 19-year follow-up study of a Japanese representative population. *Circulation Journal* 2006;70(10):1249-55.
159. Noda H, Maruyama K, Iso H, Dohi S, Terai T, Fujioka S, et al. Prediction of myocardial infarction using coronary risk scores among Japanese male workers: 3M Study. *Journal of Atherosclerosis & Thrombosis* 2010;17(5):452-9.
160. Nordestgaard BG, Adourian AS, Freiberg JJ, Guo Y, Muntendam P, Falk E. Risk factors for near-term myocardial infarction in apparently healthy men and women. *Clinical Chemistry* 2010;56(4):559-67.
161. Odell PM, Anderson KM, Kannel WB. New models for predicting cardiovascular events. *J Clin Epidemiol* 1994;47(6):583-92.
162. Onat A, Can G, Hergenc G, Ugur M, Yuksel H. Coronary disease risk prediction algorithm warranting incorporation of C-reactive protein in Turkish adults, manifesting sex difference. *Nutrition Metabolism & Cardiovascular Diseases* 2012;22(8):643-50.
163. Pencina MJ, D'Agostino RB, Sr., Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the framingham heart study. *Circulation* 2009;119(24):3078-84.
164. Petersson U, Ostgren CJ, Brudin L, Nilsson PM. A consultation-based method is equal to SCORE and an extensive laboratory-based method in predicting risk of future cardiovascular disease. *European Journal of Cardiovascular Prevention & Rehabilitation* 2009;16(5):536-40.
165. Plichart M, Celermajer DS, Zureik M, Helmer C, Jouven X, Ritchie K, et al. Carotid intima-media thickness in plaque-free site, carotid plaques and coronary heart disease risk prediction in older adults. The Three-City Study. *Atherosclerosis* 2011;219(2):917-24.
166. Pocock SJ, McCormack V, Gueyffier F, Boutitie F, Fagard RH, Boissel JP. A score for predicting risk of death from cardiovascular disease in adults with raised blood pressure, based on individual patient data from randomised controlled trials. *BMJ* 2001;323(7304):75-81.
167. Polonsky TS, McClelland RL, Jorgensen NW, Bild DE, Burke GL, Guerci AD, et al. Coronary artery calcium score and risk classification for coronary heart disease prediction. *JAMA* 2010;303(16):1610-6.
168. Prati P, Tosetto A, Casaroli M, Bignamini A, Canciani L, Bornstein N, et al. Carotid plaque morphology improves stroke risk prediction: usefulness of a new ultrasonographic score. *Cerebrovascular Diseases* 2011;31(3):300-4.
169. Qiao Q, Gao W, Laatikainen T, Vartiainen E. Layperson-oriented vs. clinical-based models for prediction of incidence of ischemic stroke: National FINRISK Study. *International Journal of Stroke* 2012;7(8):662-68.

170. Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB, Sr., et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *Lancet* 2009;373(9665):739-45.
171. Shaper AG, Pocock SJ, Phillips AN, Walker M. Identifying men at high risk of heart attacks: strategy for use in general practice. *Br Med J (Clin Res Ed)* 1986;293(6545):474-9.
172. Smith JG, Newton-Cheh C, Almgren P, Struck J, Morgenthaler NG, Bergmann A, et al. Assessment of conventional cardiovascular risk factors and multiple biomarkers for the prediction of incident heart failure and atrial fibrillation. *Journal of the American College of Cardiology* 2010;56(21):1712-9.
173. Tanabe N, Iso H, Okada K, Nakamura Y, Harada A, Ohashi Y, et al. Serum total and non-high-density lipoprotein cholesterol and the risk prediction of cardiovascular events - the JALS-ECC. *Circ J* 2010;74(7):1346-56.
174. Teramoto T, Ohashi Y, Nakaya N, Yokoyama S, Mizuno K, Nakamura H, et al. Practical risk prediction tools for coronary heart disease in mild to moderate hypercholesterolemia in Japan: originated from the MEGA study data. *Circulation Journal* 2008;72(10):1569-75.
175. Thomsen TF, Davidsen M, Ibsen H, Jorgensen T, Jensen G, Borch-Johnsen K. A new method for CHD prediction and prevention based on regional risk scores and randomized clinical trials; PRECARD and the Copenhagen Risk Score. *J Cardiovasc Risk* 2001;8(5):291-7.
176. Thorsen RD, Jacobs DR, Jr., Grimm RH, Jr., Keys A, Taylor H, Blackburn H. Preventive cardiology in practice: a device for risk estimation and counseling in coronary disease. *Prev Med* 1979;8(5):548-56.
177. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. *J Chronic Dis* 1967;20(7):511-24.
178. Tsang TS, Barnes ME, Gersh BJ, Takemoto Y, Rosales AG, Bailey KR, et al. Prediction of risk for first age-related cardiovascular events in an elderly population: the incremental value of echocardiography. *J Am Coll Cardiol* 2003;42(7):1199-205.
179. Tunstall-Pedoe H. The Dundee coronary risk-disk for management of change in risk factors. *BMJ* 1991;303(6805):744-7.
180. Vergnaud AC, Bertrais S, Galan P, Hercberg S, Czernichow S. Ten-year risk prediction in French men using the Framingham coronary score: results from the national SU.VI.MAX cohort. *Preventive Medicine* 2008;47(1):61-5.
181. Voss R, Cullen P, Schulte H, Assmann G. Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Munster Study (PROCAM) using neural networks. *Int J Epidemiol* 2002;31(6):1253-62; discussion 62-64.
182. Stork S, Feelders RA, van den Beld AW, Steyerberg EW, Savelkoul HFJ, Lamberts SWJ, et al. Prediction of mortality risk in the elderly. *American Journal of Medicine* 2006;119(6):519-25.

183. Wilson PW, Castelli WP, Kannel WB. Coronary risk prediction in adults (the Framingham Heart Study). *Am J Cardiol* 1987;59(14):91G-94G.
184. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97(18):1837-47.
185. Asselbergs FW, Hillege HL, van Gilst WH. Framingham score and microalbuminuria: combined future targets for primary prevention? *Kidney International - Supplement* 2004(92):S111-4.
186. Baxi NS, Jackson JL, Ritter J, Sessums LL. How well do the Framingham risk factors correlate with diagnoses of ischemic heart disease and cerebrovascular disease in a military beneficiary cohort? *Military Medicine* 2011;176(4):408-13.
187. Buitrago F, Calvo-Hueros JI, Canon-Barroso L, Pozuelos-Estrada G, Molina-Martinez L, Espigares-Arroyo M, et al. Original and REGICOR Framingham functions in a nondiabetic population of a Spanish health care center: a validation study. *Annals of Family Medicine* 2011;9(5):431-8.
188. D'Agostino RB, Sr., Grundy S, Sullivan LM, Wilson P. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA* 2001;286(2):180-7.
189. Drawz PE, Baraniuk S, Davis BR, Brown CD, Colon PJ, Sr., Cujyet AB, et al. Cardiovascular risk assessment: addition of CKD and race to the Framingham equation. *American Heart Journal* 2012;164(6):925-31.e2.
190. Hsia J, Rodabough RJ, Manson JE, Liu S, Freiberg MS, Graettinger W, et al. Evaluation of the American Heart Association cardiovascular disease prevention guideline for women. *Circulation. Cardiovascular Quality & Outcomes* 2010;3(2):128-34.
191. Lloyd-Jones DM, Wilson PWF, Larson MG, Beiser A, Leip EP, D'Agostino RB, et al. Framingham risk score and prediction of lifetime risk for coronary heart disease. *American Journal of Cardiology* 2004;94(1):20-4.
192. Marrugat J, D'Agostino R, Sullivan L, Elosua R, Wilson P, Ordovas J, et al. An adaptation of the Framingham coronary heart disease risk function to European Mediterranean areas. *J Epidemiol Community Health* 2003;57(8):634-8.
193. Marrugat J, Solanas P, D'Agostino R, Sullivan L, Ordovas J, Cordon F, et al. Coronary risk estimation in Spain using a calibrated Framingham function. *Rev Esp Cardiol* 2003;56(3):253-61.
194. Marrugat J, Subirana I, Comin E, Cabezas C, Vila J, Elosua R, et al. Validity of an adaptation of the Framingham cardiovascular risk function: The VERIFICA study. *Journal of Epidemiology and Community Health* 2007;61(1):40-47.
195. Park Y, Lim J, Lee J, Kim SG. Erythrocyte fatty acid profiles can predict acute non-fatal myocardial infarction. *Br J Nutr* 2009;102(9):1355-61.

196. Polak JF, Pencina MJ, Pencina KM, O'Donnell CJ, Wolf PA, D'Agostino RB, Sr. Carotid-wall intima-media thickness and cardiovascular events. *New England Journal of Medicine* 2011;365(3):213-21.
197. Rana JS, Cote M, Despres JP, Sandhu MS, Talmud PJ, Ninio E, et al. Inflammatory biomarkers and the prediction of coronary events among people at intermediate risk: the EPIC-Norfolk prospective population study. *Heart* 2009;95(20):1682-7.
198. Reissigova J, Zvarova J. The Framingham risk function underestimated absolute coronary heart disease risk in Czech men.[Erratum appears in Methods Inf Med. 2007;46(1):III]. *Methods of Information in Medicine* 2007;46(1):43-9.
199. Rodondi N, Locatelli I, Aujesky D, Butler J, Vittinghoff E, Simonsick E, et al. Framingham risk score and alternatives for prediction of coronary heart disease in older adults. *PLoS ONE [Electronic Resource]* 2012;7(3):e34287.
200. Simmons RK, Sharp S, Boekholdt SM, Sargeant LA, Khaw K-T, Wareham NJ, 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? *Archives of Internal Medicine* 2008;168(11):1209-16.
201. Stern MP, Williams K, Gonzalez-Villalpando C, Hunt KJ, Haffner SM. Does the metabolic syndrome improve identification of individuals at risk of type 2 diabetes and/or cardiovascular disease?.[Erratum appears in Diabetes Care. 2005 Jan;28(1):238]. *Diabetes Care* 2004;27(11):2676-81.
202. Suka M, Sugimori H, Yoshida K. Application of the updated Framingham risk score to Japanese men. *Hypertens Res* 2001;24(6):685-9.
203. Tohidi M, Hadaegh F, Harati H, Azizi F. C-reactive protein in risk prediction of cardiovascular outcomes: Tehran Lipid and Glucose Study. *International Journal of Cardiology* 2009;132(3):369-74.
204. Vaidya D, Yanek LR, Moy TF, Pearson TA, Becker LC, Becker DM. Incidence of coronary artery disease in siblings of patients with premature coronary artery disease: 10 years of follow-up. *American Journal of Cardiology* 2007;100(9):1410-5.
205. Weiner DE, Tighiouart H, Griffith JL, Elsayed E, Levey AS, Salem DN, et al. Kidney disease, Framingham risk scores, and cardiac and mortality outcomes. *American Journal of Medicine* 2007;120(6):552.e1-8.
206. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham Study. *Stroke* 1991;22(3):312-8.
207. Poels MMF, Steyerberg EW, Wieberdink RG, Hofman A, Koudstaal PJ, Ikram MA, et al. Assessment of cerebral small vessel disease predicts individual stroke risk. *Journal of Neurology, Neurosurgery & Psychiatry* 2012;83(12):1174-9.
208. Truelsen T, Lindenstrom E, Boysen G. Comparison of probability of stroke between the Copenhagen City Heart Study and the Framingham Study. *Stroke* 1994;25(4):802-7.

209. Wu Y, Liu X, Li X, Li Y, Zhao L, Chen Z, et al. Estimation of 10-year risk of fatal and nonfatal ischemic cardiovascular diseases in Chinese adults. *Circulation* 2006;114(21):2217-25.
210. Wu Y, Zhang L, Yuan X, Wu Y, Yi D. Quantifying links between stroke and risk factors: a study on individual health risk appraisal of stroke in a community of Chongqing. *Neurological Sciences* 2011;32(2):211-9.
211. Yip YB, Wong TKS, Chung JWY, Ko SKK, Sit JWH, Chan TMF. Cardiovascular disease: application of a composite risk index from the Telehealth System in a district community. *Public Health Nursing* 2004;21(6):524-32.
212. Diverse Populations Collaborative Group. Prediction of mortality from coronary heart disease among diverse populations: is there a common predictive function? *Heart* 2002;88(3):222-8.

Chapter 4

Performance of the Framingham risk models and Pooled Cohort Equations for predicting 10-year risk of cardiovascular disease: a systematic review and meta-analysis

Johanna AAG Damen
Romin Pajouheshnia
Pauline Heus
Karel GM Moons
Johannes B Reitsma
Rob JPM Scholten
Lotty Hooft
Thomas PA Debray

Submitted

Abstract

Background: The Framingham risk models and Pooled Cohort Equations (PCE) are widely used and advocated in guidelines for predicting the 10-year risk of developing coronary heart disease (CHD) and cardiovascular disease (CVD), respectively, in the general population. Over the past few decades, these models have been extensively validated within different populations. Our objective is to systematically review and summarize the predictive performance of three widely advocated cardiovascular risk prediction models (Framingham Wilson 1998, Framingham ATP III 2002 and PCE 2013) in men and women separately, and to assess the generalizability of performance across different subgroups and geographical regions and determine sources of heterogeneity in the findings across studies.

Methods: A search was performed in October 2017, to identify studies investigating the predictive performance of the aforementioned models. Studies were included if they externally validated one or more of the original models in the general population for men and women for the same outcome as the original model. We assessed risk of bias for each validation and extracted data on population characteristics and model performance. Performance estimates (observed expected (OE) ratio and c-statistic) were summarized using random effects models and sources of heterogeneity were explored with meta-regression.

Results: The search identified 1585 studies, of which 38 were included, describing a total of 112 external validations. Results indicate that, on average, all three models overestimate the 10-year risk of CHD and CVD (pooled OE ratio ranged from 0.58 (95% CI 0.43-0.73; Wilson men) to 0.79 (95% CI 0.60-0.97; ATP III women)). Overestimation was most pronounced for high-risk individuals, and European populations. Further, discriminative performance was better in women for all models. There was considerable heterogeneity in the c-statistic between studies, likely due to differences in eligibility criteria and population characteristics.

Conclusions: The Framingham Wilson, Framingham ATP III and PCE discriminate comparably well but all overestimate the risk of developing CVD, especially in higher risk populations. Because the extent of miscalibration substantially varied across settings, we highly recommend that researchers further explore reasons for overprediction and that the models be updated for specific populations before using them in clinical practice.

Introduction

Cardiovascular disease (CVD) is a major health burden, accounting for 17.5 million deaths worldwide in 2012.[1] Various strategies, ranging from lifestyle advice to the use of blood pressure or lipid-lowering drugs, are currently being used for timely prevention of CVD.[2-4] To effectively and efficiently implement these preventive measures, early identification of high risk individuals for targeted intervention using so-called CVD risk prediction models or risk scores is widely advocated.[5] Evidently, it is crucial that CVD risk predictions made by these models are sufficiently accurate. Inappropriate risk based management may lead to overtreatment or undertreatment, resulting in either unnecessary costs or disease burden that could have been prevented if risks were accurately predicted.

Clinical guidelines from the National Cholesterol Education Program previously advised using the Framingham Adult Treatment Panel (ATP) III model.[6] Currently, the American College of Cardiology and American Heart Association (AHA) jointly developed and advocated the Pooled Cohort Equations (PCE) to predict 10-year risk of CVD for all individuals 40 years or older.[5] Interestingly, the Framingham Wilson model[7] is, to our best knowledge, not mentioned in clinical guidelines, although it is the model that has been most extensively studied in the field of CVD risk prediction.[8]

All three models have been externally validated numerous times across different populations, and most studies showed predicted risks are overestimated (i.e. poor calibration, see box).[9-12] Some reports have, however, presented contrasting results and conclusions showing adequate calibration for these same models.[13,14]

Despite the heterogeneity found between the results and conclusions of these external validation studies, a comprehensive systematic overview and meta-analysis of all existing evidence on the predictive performance of the Framingham Wilson, ATP III, and PCE models has not yet been performed. Such evidence syntheses have become a vital tool in the cycle of prediction model development, validation and updating[15] and clearly help researchers, policy makers and clinicians to evaluate which models can be advocated in guidelines for use in daily practice. Although Framingham Wilson is not mentioned in clinical guidelines, it is relevant to review this prediction model, since many studies in the field of CVD risk prediction have externally validated this prediction model, and have used it to assess the incremental value of new predictors, or for comparison with newly developed prediction models.[8] Preferably, a meta-analysis of the performance of a prediction model should be performed to quantify the performance and to investigate sources of heterogeneity, to better understand how the model can be used in clinical practice.

We, therefore, compared the predictive performance of the Framingham Wilson, Framingham ATP III, and PCE models (see Supplement 1 for details on these prediction models and our review question). We conducted a systematic review, including critical appraisal, of all published studies that externally validated one or more of these three

models, followed by a formal meta-analysis to summarize and compare the overall predictive performance of these models, and the predictive performance across pre-defined subgroups. We explicitly did not intend to review all existing CVD risk prediction models but focused on these three most widely advocated and used models in the United States.

Box: Terminology

	Definition
Case-mix / patient spectrum	Characteristics of the study population (e.g. age, gender distribution).
Prediction horizon	Time frame in which the model predicts the outcome (e.g. predicting 10-year risk of developing a CVD event).
External validation	Estimating the predictive performance of an existing prediction model in a dataset or study population other than the dataset from which the model was developed.
Predictive performance	Accuracy of the predictions made by a prediction model, often expressed in terms of discrimination or calibration.
Discrimination	Ability of the model to distinguish between people who did and did not develop the event of interest, often quantified by the c-statistic.
Concordance (c)-statistic	Statistic that quantifies the chance that for any two individuals of which one developed the outcome and the other did not, the former has a higher predicted probability according to the model than the latter. A c-statistic of 1 means perfect discriminative ability, whereas a model with a c-statistic of 0.5 is not better than flipping a coin.[16]
Calibration	Agreement between observed event risks and event risks predicted by the model.
Observed Expected (OE) ratio	The ratio of the total number of outcome events that occurred (e.g. in 10 years) and the total number of events predicted by the model.
Calibration slope	Measure that gives an indication of the strength of the predictor effects. The calibration slope ideally equals 1. A calibration slope <1 indicates that predictions are too extreme (low risk individuals have a predicted risk that is too low, and high risk individuals are given a predicted risk that is too high). Conversely, a slope >1 indicates that predictions are too moderate.[17,18]
Model updating / recalibration	When externally validating a prediction model, adjusting the model to the dataset in which the model is validated, to improve the predictive performance of the model.

Updating the baseline hazard or risk	When externally validating a prediction model, adapting the original baseline hazard or intercept of the prediction model to the dataset in which the model is validated. This updating method corrects for differences in observed outcome incidence between the original development and external validation dataset.
Updating the common slope	When externally validating a prediction model, adapting the beta coefficients of the model using a single correction factor, to proportionally adjust for changes in predictor outcome associations.[19]
Model revision	Taking the predictors of an existing, previously developed model and fitting these in the external dataset by estimating the new predictor-outcome associations (e.g. regression coefficients).

Methods

We conducted our review based on the steps described in the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS)[20] and in a recently published guidance paper on the systematic review and meta-analysis of prediction models.[15]

Search and selection

We started with studies published before June 2013 that were already identified in two previously published systematic reviews.[8,21] Studies published after June 2013 were identified according to the following strategy. First, a search was performed in MEDLINE and Embase (October 25, 2017, Supplement 2.1.1). In addition, a citation search in Scopus and Web of Science was performed to find all studies published between 2013 and 2017 that cited the studies in which the development of one of the original models was described (Supplement 2.1.2). All studies that were identified both by the search in MEDLINE and Embase, and the citation search were screened for eligibility, first on title and abstract by one reviewer and subsequently on full text by two independent reviewers. Disagreements were solved in group discussions. The reference lists of systematic reviews identified by our search were screened to identify additional studies.

Eligibility criteria

Studies were eligible for inclusion if they described the external validation of Framingham Wilson 1998,[7] Framingham ATP III 2002,[6] and/or PCE 2013.[22] Studies were included if they externally validated these models for fatal or nonfatal coronary heart disease (CHD) in the case of Framingham Wilson and ATP III, and hard atherosclerotic CVD (here referred to as fatal or nonfatal CVD) in the case of PCE, separately for men and women, in a

general (unselected) population setting. Studies regarding specific patient populations (e.g. patients with diabetes) were excluded. Studies in which the model was updated or altered (e.g. recalibration or model revision,[23,24] see Box) before external validation were excluded if they did not provide any information on the original model's performance. Studies in which the models for men and women were combined in one validation (with one performance measure reported for men and women together instead of two separate performance measures) were excluded. Studies that assessed the incremental value of an additional predictor on top of the original model were also excluded, unless the authors explicitly reported on the external validity of the original model before adding the extra predictor. When a study population was used multiple times to validate the same model (i.e. multiple publications describing a certain study cohort), the external validation with eligibility criteria and predicted outcome that most closely resembled our review question (Supplement 1.1) was included, to avoid introducing bias because of duplicate data.[25]

Data extraction and critical appraisal

For each included study, data were extracted on study design, population characteristics, participant enrolment, study dates, prediction horizon, predicted outcomes, predictors, sample size, model updating methods, and model performance (Supplement 2.2). Risk of bias was assessed based on a combination of the CHARMS checklist[20] and a preliminary version of the Cochrane Prediction study Risk Of Bias Assessment Tool (PROBAST).[26,27] Risk of bias was assessed for each validation, across five domains: participant selection (e.g. study design, in- and exclusions), predictors (e.g. differences in predictor definitions), outcome (e.g. same definition and assessment for every participant), sample size and participant flow (e.g. handling of missing data), analyses (e.g. handling of censoring). After several rounds of piloting and adjusting the data extraction form in a team of three reviewers, data were extracted by one of the three reviewers. Risk of bias was independently assessed by pairs of reviewers. Disagreements were solved after discussion or by a third reviewer.

Information was extracted on model discrimination and calibration, before and, if reported, after model updating, in terms of the reported concordance (c)-statistic and total observed versus expected (OE) ratio. If relevant information was missing (e.g. standard error of performance measure or population characteristics), we contacted the authors of the corresponding study. If no additional information could be obtained, we approximated missing information using formulas described by Debray et al.[15] (Supplement 2.3). If reported, calibration was also extracted for different risk categories. If the OE ratio was reported for shorter time intervals (e.g. 5 years) we extrapolated this to 10 years assuming a Poisson distribution (Supplement 2.3).

Statistical analyses

We performed meta-analyses of the 10-years total OE ratio and the c-statistic. Based on previous recommendations,[15,28] we pooled the log OE ratio and logit c-statistic using random-effects meta-analysis. Further, we stratified the meta-analysis by model and gender, resulting in six main groups: Wilson men, Wilson women, ATP III men, ATP III women, PCE men, PCE women. We calculated 95% confidence intervals (CI) and (approximate) 95% prediction intervals (PI) to quantify uncertainty and the presence of between-study heterogeneity. The CI indicates the precision of the summary performance estimate and the PI provides boundaries on the likely performance in future model validation studies that are comparable to the studies included in the meta-analysis, and can thus be seen as an indication of model generalizability (Supplement 2.4.1).[29] The observed and predicted probabilities in risk categories were plotted against each other and combined into a summary estimate of the calibration slope using mixed effects models (Supplement 2.4.2).

Since between-study heterogeneity in estimates of predictive performance is expected due to differences in the design and execution of validation studies,[15] we investigated whether the c-statistic differed between validation studies with different eligibility criteria or actual case-mix. Furthermore, we performed univariable random effects meta-regression analyses to investigate the influence of case-mix differences (e.g. due to differences in eligibility criteria) on the OE ratio and c-statistic (Supplement 2.4.3). Several pre-specified sensitivity analyses were performed in which we studied the influence of risk of bias and alternative weighting methods in the meta-analysis on our findings (Supplement 2.4.4). All analyses were performed in R version 3.3.2,[30] using the packages metafor,[31] mvmeta,[32] metamisc,[33] and lme4.[34]

Results

Identification and selection of studies

We first identified 100 potentially eligible studies from previously conducted systematic reviews. An additional search identified 1585 studies since June 2013 (Figure 1). Of these 1685 studies, 304 studies were screened on full-text and data were extracted for 61 studies, describing 167 validations of the performance of one or more of the three models. Finally, 38 studies (112 validations) met our eligibility criteria.

Description of included validations

In 112 validations (Supplement 3.3), the Framingham Wilson model was validated 38 times (men: 23, women: 15), Framingham ATP III 13 times (men: 7, women: 6), and PCE 61 times (men: 30, women: 31). Study participants were recruited between 1965 and 2008, and originated from North America (56), Europe (29), Asia (25), and Australia (2). We excluded 18 and 9 external validations because the OE ratio and c-statistic, respectively, were not available, and subsequently excluded 20 and 26 external validations for the OE ratio and c-statistic, respectively, because cohorts were used multiple times to validate the same model. This resulted in the inclusion of 74 validations in the analyses of the OE ratio and 77 validations in the analyses of the c-statistic (Figure 1).

Risk of bias

For participant selection, most validations scored low risk of bias (n=60 (81%) and n=64 (83%) for validations reporting OE ratio and c-statistic, respectively. Figure 2). Risk of bias for predictors was often unclear (n=22 (30%) and n=24 (31%), for OE ratio and c-statistic), due to poor reporting of predictor definitions and measurement methods. Most validations scored low risk of bias on outcome (n=53 (72%), n=59 (77%)). More than three quarters of the validations scored high risk of bias for sample size and participant flow (n=59 (80%) and n=60 (78%)), often due to inadequate handling of missing data (i.e. simply ignoring). Low risk of bias for analysis was scored in 51 (70%) and 50 (65%) validations, for OE ratio and c-statistic respectively. In total, 62 (84%) and 63 (82%) validations scored high risk of bias for at least one domain, and 4 (5%) and 6 (8%) validations scored low risk of bias for all five domains, for OE ratio and c-statistic, respectively.

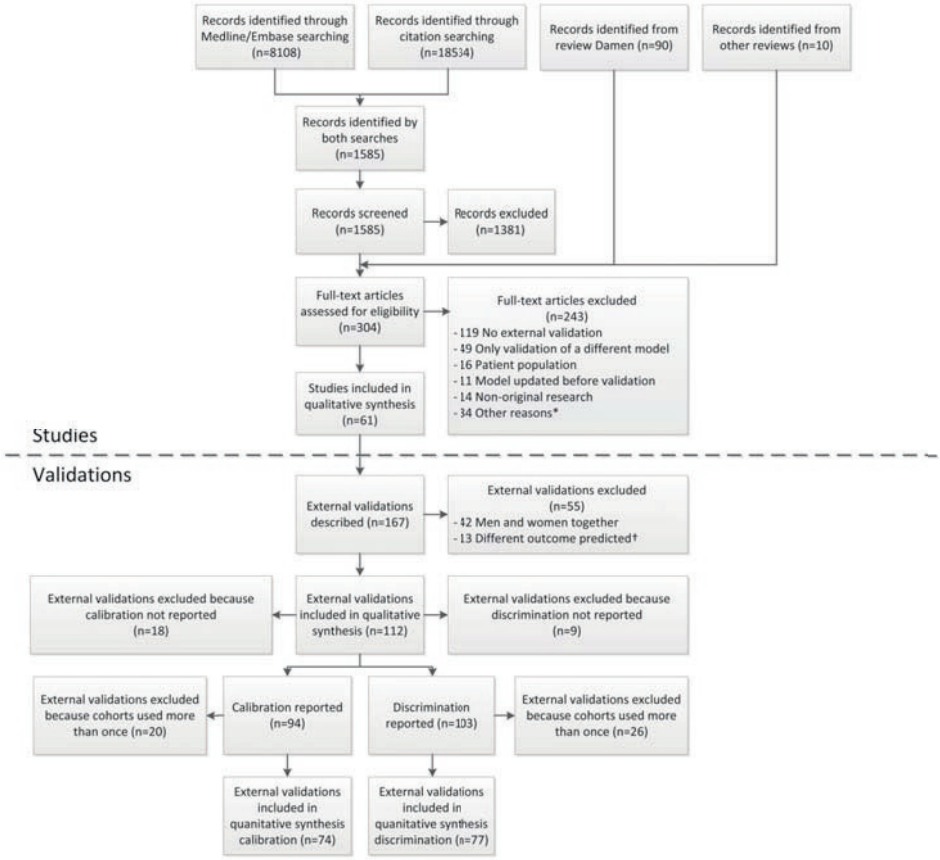


Figure 1: Flow diagram of selected studies. Two searches were performed; one in MEDLINE and Embase and one in Scopus and Web of Science. Only studies identified by both searches were screened for eligibility, supplemented with records identified from previous systematic reviews. One study could describe more than one external validation (e.g. one for men and one for women) therefore, 61 studies described 167 external validations. Calibration was available for 94 validations (41 directly reported, 19 provided by the authors on request, 34 estimated from calibration tables and calibration plots), and discrimination for 103 validations (91 c-statistics directly reported, 12 provided by the authors on request. Precision of c-statistic: 45 directly reported, 24 provided by the authors, 32 estimated from the sample size, and 2 not reported). Some external validations were excluded because cohorts were used more than once to validate the same model (Supplement 3.2). *E.g. no cardiovascular outcome, not written in English. †The Framingham Wilson and ATP III models were developed to predict the risk of fatal or nonfatal coronary heart disease and the PCE model was developed to predict the risk of fatal or nonfatal cardiovascular disease. External validations that used a different outcome were excluded from the analyses (Supplement 3.1).

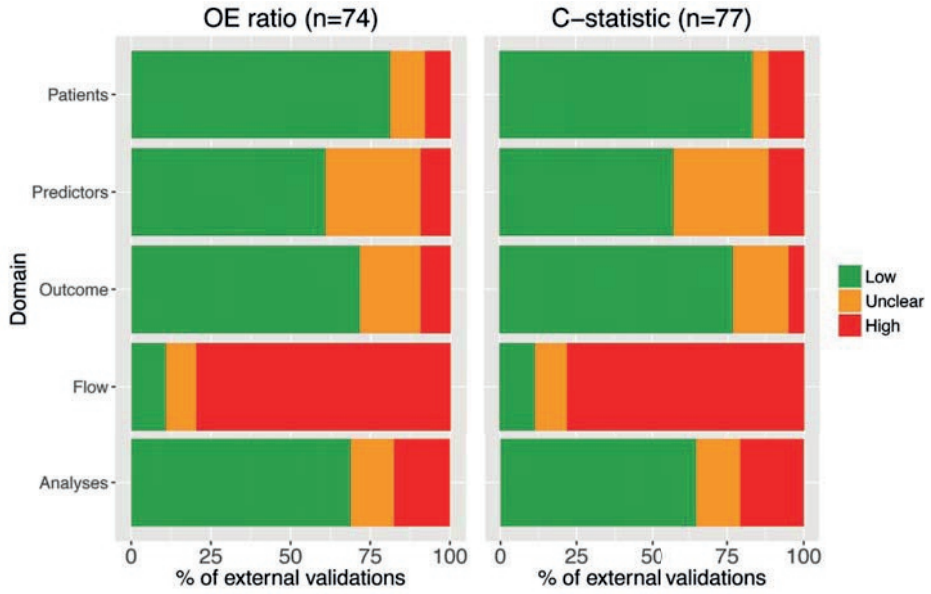


Figure 2: Summary of risk of bias assessments for validations included in the meta-analyses of OE ratio (74 validations) and c-statistic (77 validations).

Calibration

Figure 3 shows the calibration of the six main models, as depicted by their 10-year total OE ratio. For 24 out of 74 validations (32%), maximum follow-up was shorter than 10 years. For 20 out of these 24 (83%), information was available to extrapolate the OE ratio to 10 years. Most studies showed overprediction, indicating that 10-year risk predictions provided by the models were typically higher than observed in the validation datasets. For the Wilson model, the number of events predicted by the model was lower than the actual number of events in two studies (one in healthy siblings of patients with premature coronary artery disease,[35] and one in community-dwelling individuals aged 70–79[36]). For PCE, underestimation of the number of events occurred in Chinese[37] and Korean[38] populations.

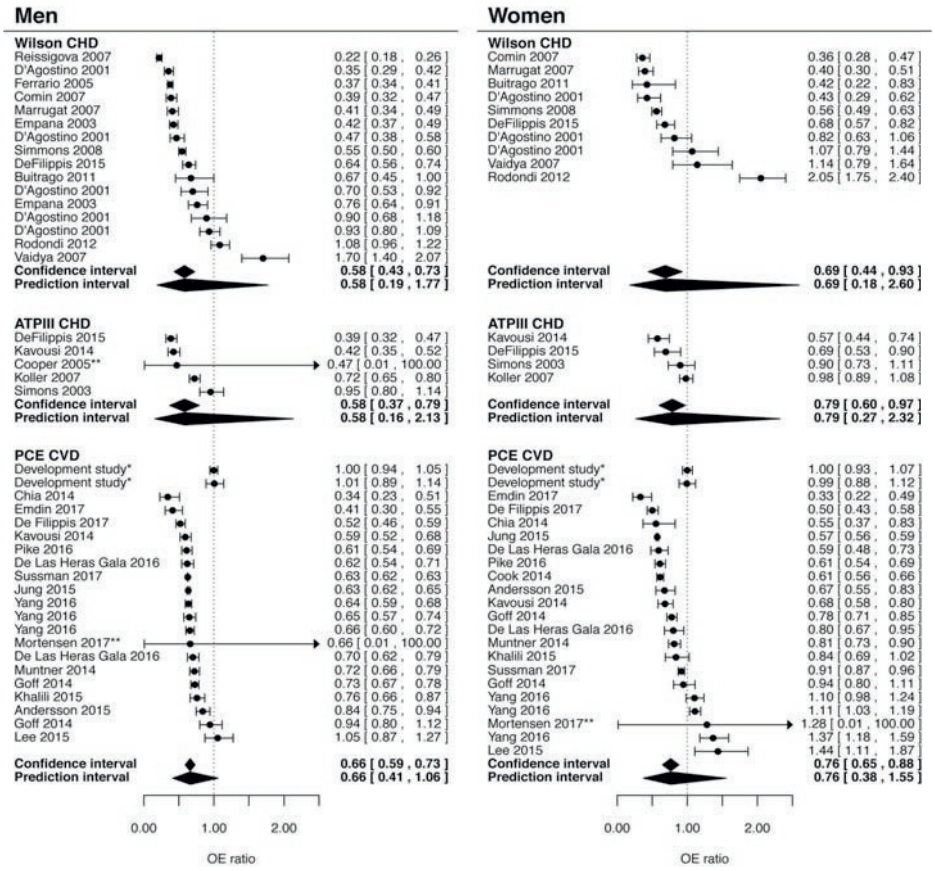


Figure 3: Meta-analysis of the OE ratio in external validations, with 95% confidence intervals and 95% prediction intervals per model. The performance of the model in the development study is shown in the first rows (only reported for PCE). This estimate is not included in calculating the pooled estimate of performance. *Performance of the model in the development population after internal validation. The first row contains the performance of the model for Whites, the second for African Americans. **Standard error was not available. CHD: Coronary heart disease, CVD: cardiovascular disease.

Meta-analysis revealed a considerable degree of between-study heterogeneity in OE ratios (Figure 3), but with clear overprediction, as summary OE ratios ranged from 0.58 (Wilson men and ATP III men) to 0.79 (ATP III women). Additional analyses revealed that overprediction is more pronounced in high-risk patients, for all models (Figure 4). The results of the summary calibration slope suggest that miscalibration of the Framingham Wilson and ATP III models, and PCE men model was mostly related to heterogeneity in baseline risk (as the summary calibration slope is close to 1), while for PCE women we found a slope around 0.8, suggesting that this model was overfitted or does not transport well to new populations (Supplement 3.4). For 38 validations the model was subsequently updated, of which 24 reported the OE ratio after updating. The OE ratio improved after updating (0.65 (IQR 0.46-0.86) before vs. 0.84 (IQR 0.70-0.91) after updating).

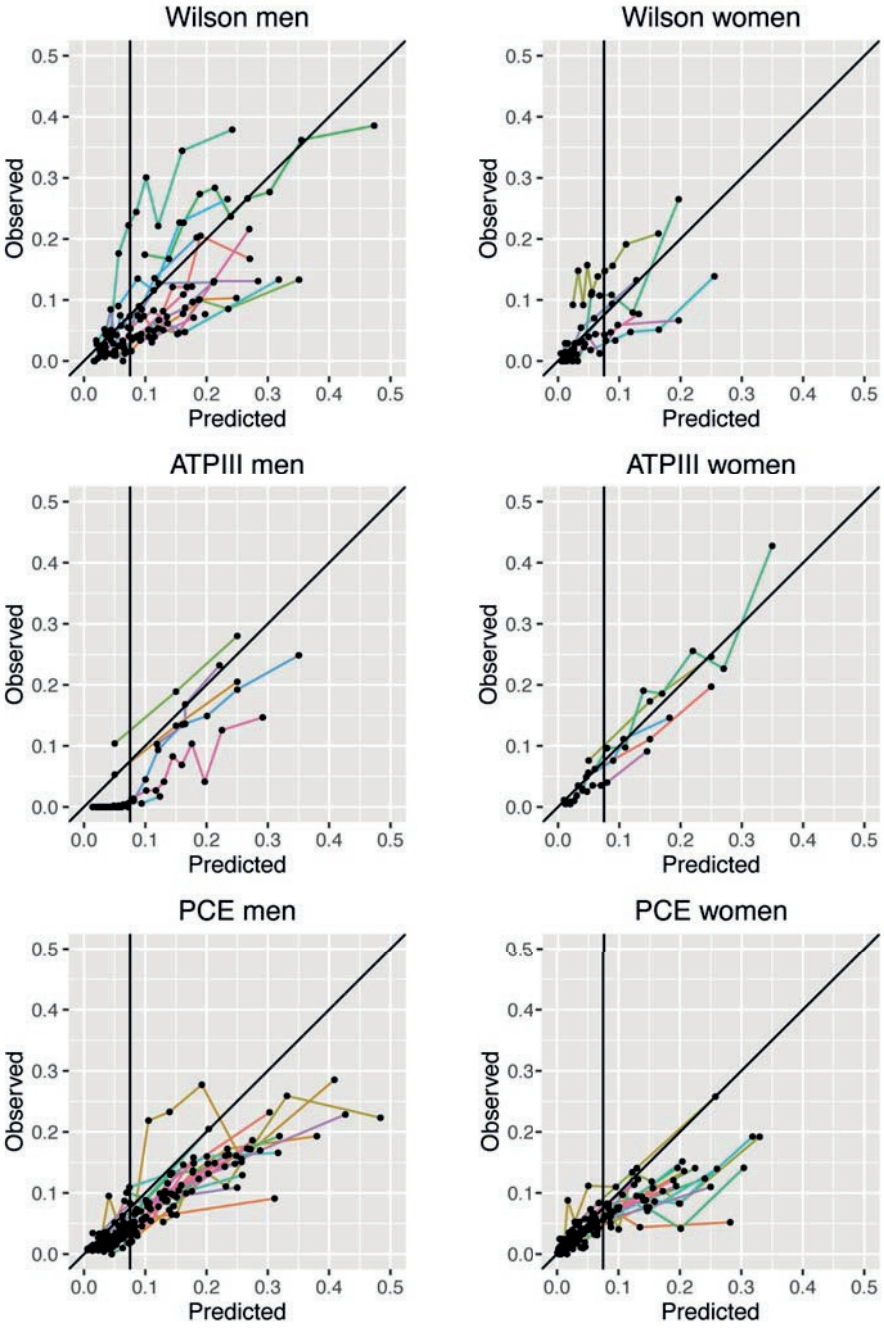


Figure 4: Calibration plots of the Framingham Wilson, ATP III and PCE models. Each line represents one external validation. The diagonal line represents perfect agreement between observed and predicted risks. All points below that line indicate that more events were predicted than observed (overprediction) and points above the line indicate fewer events were predicted than observed (underprediction). The vertical black line represents a treatment threshold of 7.5%.[39] CI: confidence interval, PI: prediction interval.

Discrimination

For all models, discriminative performance was slightly better for women than for men, although there was considerable variation between studies (Figure 5). For 40 out of 74 validations model updating was performed, of which 13 reported the c-statistic after update. Results indicate that the c-statistic did not change after updating (median 0.71 (IQR 0.66-0.72) before vs. 0.72 (IQR 0.69-0.76) after update)

Sensitivity analyses

Sensitivity analyses revealed no effect of study quality and different weighting strategies on the pooled performance of the models, both for calibration and discrimination (Supplement 3.5).

Factors that influence performance of the models

For women, the highest c-statistics were reported in studies with large variety in case-mix. For men, such a trend was not visible (Figure 6). The OE ratio for the Wilson model in the United States was closer to 1 compared to Europe, but the number of external validations per subgroup was very small (Supplement 3.6.1). Furthermore, the OE ratio appeared to decrease (further away from 1, i.e. more overprediction) with increasing mean total cholesterol. No evidence was found of an association between the OE ratio and other case-mix variables or start date of participant recruitment. The c-statistic appeared to decrease with increasing mean age, mean systolic blood pressure and standard deviation of HDL cholesterol, and to increase with increasing standard deviation of age and total cholesterol (Supplement 3.6.2). No statistically significant associations were found between the c-statistic and other variables.

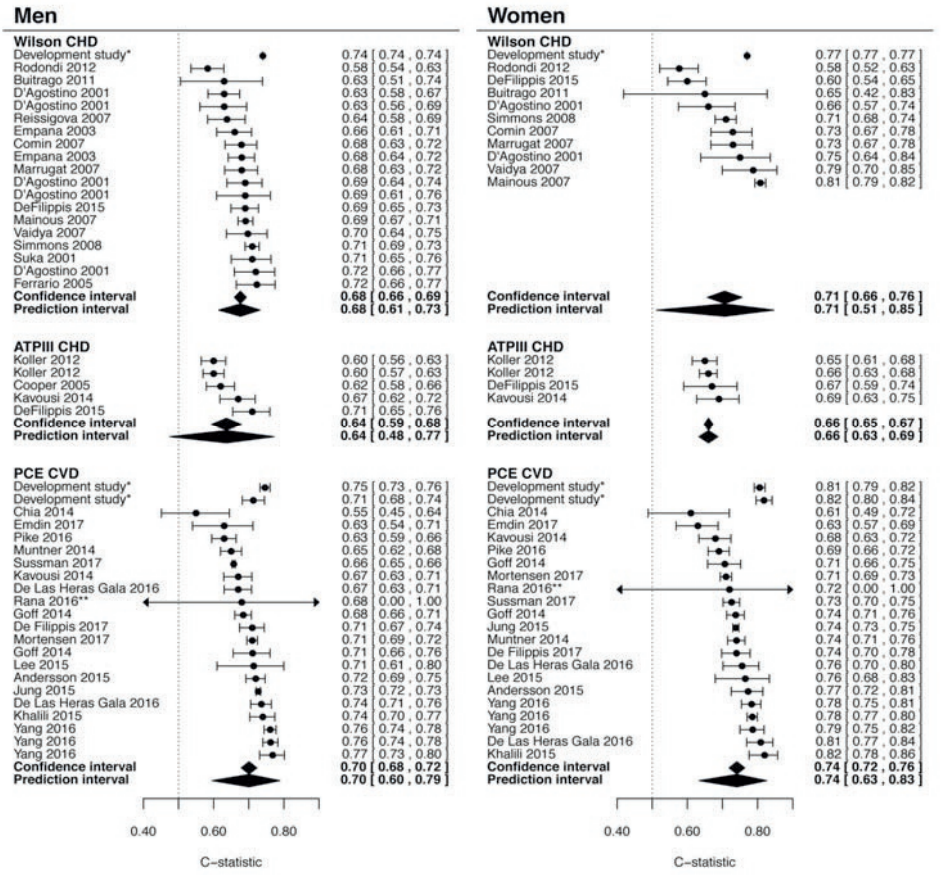
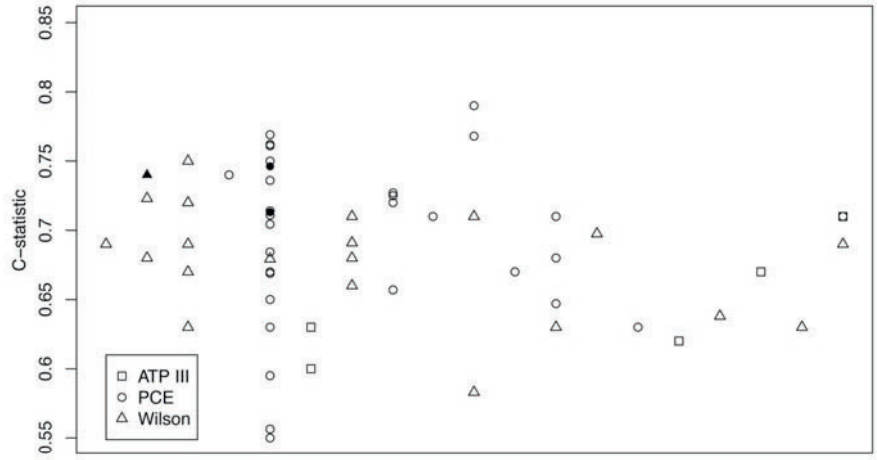


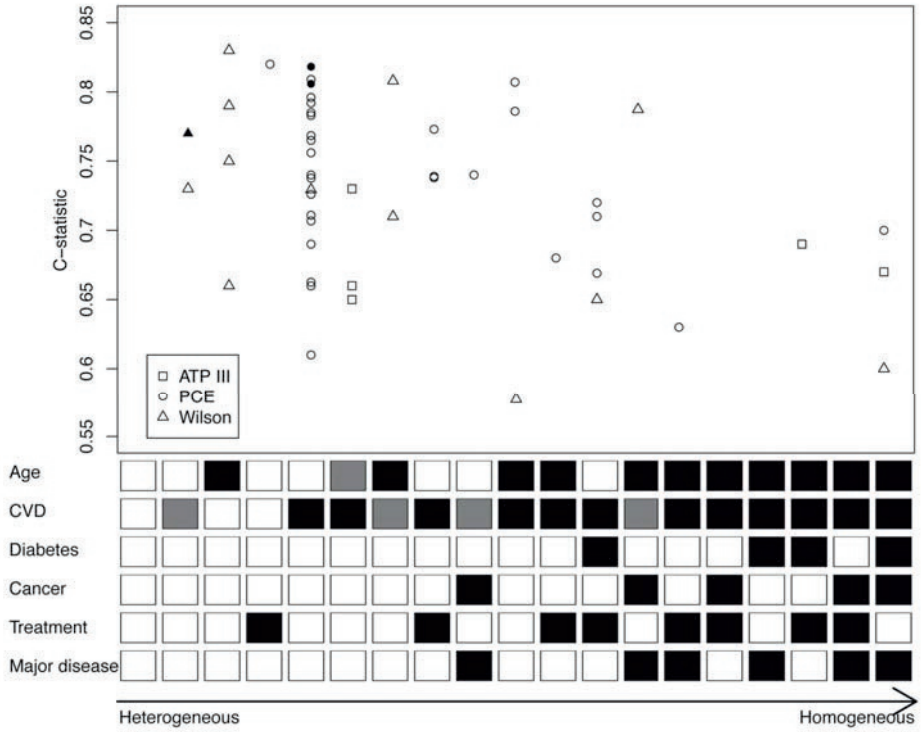
Figure 5: C-statistic in external validations, with 95% confidence intervals and 95% prediction intervals per model. The performance of the model in the development study is shown in the first row(s) (not reported for ATP III) and is not included in the pooled estimate of performance. *Performance of the model in the development population (Wilson (no standard error reported)) and after 10x10 cross-validation (PCE). For PCE, the first row contains the performance of the White model, the second the African American model. **Standard error was not available. CHD: coronary heart disease, CVD: cardiovascular disease.

Figure 6: (right page) C-statistic for different combinations of eligibility criteria. The open squares, circles and triangles represent validations of the ATP III, PCE and Wilson model, respectively. The black circles and triangles represent the performance of the PCE models for Whites and African-Americans, and Wilson models, in the development populations. Lower part: for age, white means a broad age range was included (difference between upper and lower age limit >30 years), black means a narrow age range was included (difference between upper and lower age limit ≤30 years), and grey means age was not reported. For CVD, white means no exclusion of people with CHD or CVD, grey means people with previous CHD events were excluded from the study, and black means people with previous CVD events were excluded from the study. For diabetes, cancer, major disease, white means that no restrictions were reported and black means that people with these conditions were excluded. For treatment, white means no restrictions and black means people who were receiving any treatment to lower their risk of CVD (e.g. antihypertensives) were excluded from the study.

Men


C–statistic
0.55
0.6
0.65
0.7
0.75
0.8
0.85
ATP III
PCE
Wilson

Women


C–statistic
0.55
0.6
0.65
0.7
0.75
0.8
0.85
ATP III
PCE
Wilson
Age
CVD
Diabetes
Cancer
Treatment
Major disease
Heterogeneous
Homogeneous

Discussion

Summary of findings

We systematically reviewed the performance of the Framingham Wilson, Framingham ATP III, and PCE models for predicting 10-year risk of CHD or CVD for men and women separately in the general population. We found only small differences in pooled performance between the three models, but large differences in performance between validations of the same model. Although we mostly had to rely on indirect comparisons of the models, we found that performance of all three models was consistently better in women than in men for both discrimination and calibration. This can probably be attributed to a stronger association between risk factors and CVD in women compared to men.[40] In agreement with previous systematic reviews,[21,41-43] we found that all models overestimated the risk of CHD or CVD, and this overestimation was more pronounced in European populations compared to the United States. Overprediction clearly declined when the validated models were adjusted (e.g. via updating the baseline hazard) to the validation setting at hand. This indicates that the prediction models should not simply be advocated or applied in guidelines or clinical practice, but first tailored to the setting in which they are to be applied. Although it was not possible to identify statistically significant sources of heterogeneity, we found that discriminative performance tends to increase as populations become more diverse, i.e. with a wider case-mix. This effect has previously been explained.[44-46]

Reasons for overprediction

There could be several reasons for the observed overprediction, which have also extensively been discussed previously with regards to the PCE.[43,47,48] First, differences in eligibility criteria (e.g. the exclusion of participants with previous CVD events) across validation studies may have affected calibration. Second, the three prediction models have been (partly) developed using data from the 1970s and since then treatment of people at high risk for a CVD event has changed considerably, such as the introduction of statins in 1987.[49] The increased use of effective treatments over time aimed at preventing CVD events will lower the observed number of events in more recent validation studies, resulting in overestimation of risk in these validation populations.[50-52] This would also explain why overprediction was most pronounced in high-risk individuals and why we found more overprediction in studies with increasing mean total cholesterol levels. We hypothesized that the degree of overprediction would increase over the years,[21,41] however this could not be confirmed statistically. About one third of validations of the PCE excluded participants receiving treatment to lower CVD risk at baseline, but we found no difference in performance between validations that did or did not exclude these participants. However, as the use of risk-lowering medication during follow-up was rarely reported in these studies, we cannot rule out an effect of incident treatment

use on model performance.[52] Third, we found more overestimation of risk in European populations compared to those of the United States whereas in some Asian populations an underestimation was seen. Both suggest that differences between these populations in, for example, unmeasured CVD risk factors and in the use of preventive CVD strategies (e.g. medical treatment or lifestyle programs), are responsible. Following the recently issued Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guideline,[53,54] and the guidance on adjusting for treatment use in prediction model studies,[51,52,55] we also strongly recommend investigators of future prediction model studies to record the use of treatment during follow-up. Finally, rather than overprediction by the models, there could also be issues in the design of the external validation studies that give rise to a lower number of identified events. Underascertainment or misclassification of outcome events, unusually high rates of people receiving treatment, short follow-up duration, and inclusion of ethnicities not included in development of the models, have been mentioned as reasons for the overprediction we observe.[56-60] Others have however shown that the overestimation could not be fully explained by treatment use and missed outcome events.[50,61]

Implications for practice and research

According to the ACC-AHA guidelines,[5] risk lowering treatment is considered in people 40-75 years old, without diabetes, with LDL cholesterol levels between 70 and 189 mg/dl and 10-year predicted risk of CVD ≥ 7.5%. After a discussion between clinician and patient about adverse effects and patient preferences, it is decided whether risk lowering treatment is initiated. The observed overprediction is problematic as this might change the population eligible for risk lowering treatment. Unfortunately, this is true for all three CVD risk prediction models. As the meta-analysis indicates that overprediction does not consistently occur across different settings and populations, there is no simple solution to address this problem. From the studies that provided data on calibration in subgroups, we found that overestimation was more pronounced in high-risk individuals. When the (over)estimation of the absolute risk is already beyond the treatment probability threshold, it will not influence treatment decisions, although overestimated risk estimations might still influence the intensity (dose and frequency) of administered treatments. For people at lower risk this might, however, result in crossing the treatment probability boundary when, actually, they are at lower risk.

In general, the performance of prediction models tends to vary substantially across different settings and populations, due to differences in case-mix and health care systems.[62] Hence, one external validation may not be sufficient to claim adequate performance and multiple validations are necessary to get insight in the generalizability of prediction models.[46] Based on this review, it can be concluded that none of the models offer reliable predictions unless (at least) their baseline risk or hazard (and, if applicable, population means of the predictors in the model) are recalibrated to the local setting.

Studies that reported performance of the model before and after update showed that performance indeed improves after update.[11,13,14,38,63,64] As previously emphasized, more extensive revision methods are often not needed.[23,24,65] Hence, it appears that conventional predictors, such as age, smoking, diabetes, blood pressure and cholesterol, are still relevant indicators of 10-year CHD or CVD risk, and their association with CVD events have largely remained stable. The need for updating CVD risk prediction models has already been discussed more than 15 years ago,[14,66] but still nothing has changed. We believe this should change now, especially since nowadays applying simple model updating is becoming increasingly possible, due to improvements in the storage of the information required to update a model. A nice example of tailoring CVD risk prediction models to specific populations, is the Globorisk prediction model which can easily be tailored to different countries using country-specific data on the population prevalence of outcomes and predictors,[67] and the SCORE model which has been tailored to many European countries using national mortality statistics.[68-71]

These suggestions, however, offer no short-term solution for practitioners currently using the three reviewed prediction models. Fortunately, a systematic review has shown that the prevalence of common CVD risk factors decreases (e.g. cholesterol levels drop) in populations where CVD risk prediction models and their corresponding treatment guidance are being used.[72] Furthermore, statins have been proven effective with limited adverse events.[4] Finally, we advise practitioners to choose a model that predicts a clinically relevant outcome (for example (according to the AHA), CVD rather than only CHD, since stroke and CHD share pathophysiological mechanisms[22,73]), consists of predictors available in their situation, and is developed or updated in a setting that closely resembles their setting.

Limitations

This study has several limitations. Firstly, we focused on the three most validated and used prediction models in the United States, while in Europe many more prediction models are currently used for predicting cardiovascular risk, such as QRISK3[74] and SCORE.[68] The differences between all these models are however limited, as most models include the same core set of predictors. Therefore, we believe our results can be generalized to other prediction models. Secondly, we had to rely on what is reported by the authors of primary validation studies and we unfortunately had to exclude relevant validations from our meta-analyses because of unreported information which we could not obtain from the authors. Only 19 out of 61 authors were able to provide us with additional information and we had to exclude 9 validations for the c-statistic and 18 for the OE ratio. Thirdly, the total OE ratio, while commonly reported, only provides an overall measure of calibration. To overcome this problem, we extracted information on the OE ratio in categories of predicted risk, which showed there was more overestimation of risk in the highest categories of predicted risk. Based on this information, we calculated

the calibration slope, which suggested that miscalibration of the Framingham Wilson and ATP III models and PCE men model was mostly related to heterogeneity in baseline risk, while for PCE women the model is overfitted or does not transport well to new populations. In addition, more clinically relevant measures, such as net benefit, could not be considered in this meta-analysis due to the lack of reporting of these measures.[8] Fourthly, because of the low number of external validation studies, especially for the ATP III model, we did not perform meta-regression analyses for this model. Unfortunately, the relatively small sample size makes it difficult to draw firm conclusions on the sources of observed heterogeneity. Fifthly, the exclusion of non-English studies could have influenced the geographical representation. However, since only 1 full-text article was excluded for this reason, we believe the effect on our results is limited.

Conclusion

The Framingham Wilson, Framingham ATP III and PCE prediction models, perform equally well in predicting the risk of CHD or CVD, but there is large variation between validations. All three prediction models overestimate the risk of CHD or CVD, which could lead to overtreatment. Therefore, before advocating their use in a clinical guideline or practice, we recommend to first further investigate reasons for overprediction and subsequently tailor or recalibrate the model to the setting at hand. Investigators and guidelines should focus on offering health care professionals the right tools and information on how to tailor these existing models to their specific settings,[23,24,65] rather than providing yet another CVD risk model for another specific subpopulation.

Acknowledgements

The authors would like to acknowledge René Spijker for performing the search in MEDLINE and Embase, and Gary Collins and Doug Altman for their valuable input in designing the study and interpreting the results. Furthermore, we acknowledge all authors of included studies, who provided additional information on their studies: Dr. Andersson, Dr. Asgari, Dr. van den Brandt, Dr. Buitrago, Dr. Chamberlain, Dr. Chia, Dr. Cook, Dr. DeFilippis, Dr. Ferrario, Dr. Giovanni, Dr. Hadaegh, Dr. van der Heijden, Dr. Khalili, Dr. Koenig, Dr. Locatelli, Dr. Marrugat, Dr. Merry, Dr. Reissigová, Dr. Ridker, Dr. Rodondi, Dr. Schouten, Dr. Simmons, Dr. Subirana, Dr. Sussman, Dr. Tan, Dr. Vaidya, Dr. Vila, Dr. Williams, Dr. Young, Dr. Zvárová.

References

1. WHO. Cardiovascular diseases (CVDs) Fact sheet N°317, 2016.
2. Korczak D, Dietl M, Steinhauser G. Effectiveness of programmes as part of primary prevention demonstrated on the example of cardiovascular diseases and the metabolic syndrome. *GMS Health Technol Assess* 2011;7:Doc02.
3. Law MR, Morris JK, Wald NJ. Use of blood pressure lowering drugs in the prevention of cardiovascular disease: meta-analysis of 147 randomised trials in the context of expectations from prospective epidemiological studies. *BMJ* 2009;338:b1665.
4. Taylor F, Huffman MD, Macedo AF, Moore THM, Burke M, Smith GD, et al. Statins for the primary prevention of cardiovascular disease. *Cochrane Libr* 2013.
5. Stone NJ, Robinson JG, Lichtenstein AH, Bairey Merz CN, Blum CB, Eckel RH, et al. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014;129(25 Suppl 2):S1-45.
6. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* 2002;106(25):3143-421.
7. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97(18):1837-47.
8. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;353:i2416.
9. Kavousi M, Leening MJ, Nanchen D, Greenland P, Graham IM, Steyerberg EW, et al. Comparison of application of the ACC/AHA guidelines, Adult Treatment Panel III guidelines, and European Society of Cardiology guidelines for cardiovascular disease prevention in a European cohort. *JAMA* 2014;311(14):1416-23.
10. DeFilippis AP, Young R, Carrubba CJ, McEvoy JW, Budoff MJ, Blumenthal RS, et al. An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort. *Ann Intern Med* 2015;162(4):266-75.
11. Reissigova J, Zvarova J. The Framingham risk function underestimated absolute coronary heart disease risk in Czech men. *Methods Inf Med* 2007;46(1):43-9.
12. Comin E, Solanas P, Cabezas C, Subirana I, Ramos R, Gene-Badia J, et al. Estimating cardiovascular risk in Spain using different algorithms. *Rev Esp Cardiol* 2007;60(7):693-702.
13. Khalili D, Asgari S, Hadaegh F, Steyerberg EW, Rahimi K, Fahimfar N, et al. A new approach to test validity and clinical usefulness of the 2013 ACC/AHA guideline on statin therapy: A population-based study. *Int J Cardiol* 2015;184(1):587-94.

14. D'Agostino RB, Sr., Grundy S, Sullivan LM, Wilson P. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA* 2001;286(2):180-7.
15. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;356:i6460.
16. Harrell F. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*: Springer, 2015.
17. Steyerberg E. *Clinical prediction models: a practical approach to development, validation, and updating*: Springer Science & Business Media, 2008.
18. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;35(29):1925-31.
19. Su TL, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods for clinical prediction models. *Stat Methods Med Res* 2016.
20. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11(10):e1001744.
21. Beswick AD, Brindle P, Fahey T, Ebrahim S. *A Systematic Review of Risk Scoring Methods and Clinical Decision Aids Used in the Primary Prevention of Coronary Heart Disease (Supplement)*. London: Royal College of General Practitioners, 2008.
22. Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014;129(25 Suppl 2):S49-73.
23. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23(16):2567-86.
24. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;61(1):76-86.
25. Tramer MR, Reynolds DJ, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. *BMJ* 1997;315(7109):635-40.
26. Wolff R, Collins GS, Kleijnen J, Mallett S, Reitsma JB, Riley R, et al. PROBAST: a risk of bias tool for prediction modelling studies. *24th Cochrane Colloquium*. Seoul, South Korea: Cochrane Database of Systematic Reviews, 2016.
27. Ensor J, Riley RD, Moore D, Snell KI, Bayliss S, Fitzmaurice D. Systematic review of prognostic models for recurrent venous thromboembolism (VTE) post-treatment of first unprovoked VTE. *BMJ Open* 2016;6(5):e011190.

28. Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res* 2017:962280217705678.
29. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2015.
30. R: A language and environment for statistical computing [program]. Vienna, Austria: R Foundation for Statistical Computing, 2016.
31. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 2010;36(3):1-48.
32. Gasparrini A, Armstrong B, Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Stat Med* 2012;31(29):3821-39.
33. Debray TP. Metamisc: Diagnostic and Prognostic Meta-Analysis. 2017.
34. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw* 2015;67(1).
35. Vaidya D, Yanek LR, Moy TF, Pearson TA, Becker LC, Becker DM. Incidence of coronary artery disease in siblings of patients with premature coronary artery disease: 10 years of follow-up. *Am J Cardiol* 2007;100(9):1410-5.
36. Rodondi N, Locatelli I, Aujesky D, Butler J, Vittinghoff E, Simonsick E, et al. Framingham risk score and alternatives for prediction of coronary heart disease in older adults. *PLoS One* 2012;7(3):e34287.
37. Lee CH, Woo YC, Lam JKY, Fong CHY, Cheung BMY, Lam KSL, et al. Validation of the Pooled Cohort equations in a long-term cohort study of Hong Kong Chinese. *J Clin Lipidol* 2015;9(5):640-46.
38. Jung KJ, Jang Y, Oh DJ, Oh BH, Lee SH, Park SW, et al. The ACC/AHA 2013 pooled cohort equations compared to a Korean Risk Prediction Model for atherosclerotic cardiovascular disease. *Atherosclerosis* 2015;242(1):367-75.
39. Eckel RH, Jakicic JM, Ard JD, de Jesus JM, Houston Miller N, Hubbard VS, et al. 2013 AHA/ACC guideline on lifestyle management to reduce cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014;129(25 Suppl 2):S76-99.
40. Paynter NP, Everett BM, Cook NR. Cardiovascular disease risk prediction in women: is there a role for novel biomarkers? *Clin Chem* 2014;60(1):88-97.
41. Brindle P, Beswick A, Fahey T, Ebrahim S. Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. *Heart* 2006;92(12):1752-9.
42. Eichler K, Puhan MA, Steurer J, Bachmann LM. Prediction of first coronary events with the Framingham score: a systematic review. *Am Heart J* 2007;153(5):722-31, 31.e1-8.

43. Cook NR, Ridker PM. Calibration of the Pooled Cohort Equations for Atherosclerotic Cardiovascular Disease: An Update. *Ann Intern Med* 2016.
44. Knottnerus JA. Diagnostic prediction rules: principles, requirements and pitfalls. *Prim Care* 1995;22(2):341-63.
45. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010;172(8):971-80.
46. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2014.
47. Cook NR, Ridker PM. Further insight into the cardiovascular risk calculator: the roles of statins, revascularizations, and underascertainment in the Women's Health Study. *JAMA* 2014;174(12):1964-71.
48. Ridker PM, Cook NR. The Pooled Cohort Equations 3 Years On: Building a Stronger Foundation. *Circulation* 2016;134(23):1789-91.
49. Tobert JA. Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. *Nat Rev Drug Discov* 2003;2(7):517-26.
50. Cook NR, Ridker PM. Further insight into the cardiovascular risk calculator: the roles of statins, revascularizations, and underascertainment in the Women's Health Study. *JAMA Intern Med* 2014;174(12):1964-71.
51. Pajouheshnia R, Peelen LM, Moons KGM, Reitsma JB, Groenwold RHH. Accounting for treatment use when validating a prognostic model: a simulation study. *BMC Med Res Methodol* 2017;17(1):103.
52. Pajouheshnia R, Damen JA, Groenwold RH, Moons KG, Peelen LM. Treatment use in prognostic model research: a systematic review of cardiovascular prognostic studies. *Diagnostic and Prognostic Research* 2017;1(1):15.
53. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162(1):55-63.
54. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.
55. Groenwold RH, Moons KG, Pajouheshnia R, Altman DG, Collins GS, Debray TP, et al. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *J Clin Epidemiol* 2016;78:90-100.
56. Muntner P, Safford MM, Cushman M, Howard G. Comment on the reports of over-estimation of ASCVD risk using the 2013 AHA/ACC risk equation. *Circulation* 2014;129(2):266-7.

57. Muntner P, Colantonio LD, Cushman M, Goff DC, Jr., Howard G, Howard VJ, et al. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *JAMA* 2014;311(14):1406-15.
58. Krumholz HM. The new cholesterol and blood pressure guidelines: perspective on the path forward. *JAMA* 2014;311(14):1403-5.
59. Goff DC, Jr., D'Agostino RB, Sr., Pencina M, Lloyd-Jones DM. Calibration and Discrimination Among Multiple Cardiovascular Risk Scores in a Modern Multiethnic Cohort. *Ann Intern Med* 2015;163(1):68.
60. Spence JD. Statins and ischemic stroke. *JAMA* 2014;312(7):749-50.
61. Cook NR, Ridker PM. Response to Comment on the reports of over-estimation of ASCVD risk using the 2013 AHA/ACC risk equation. *Circulation* 2014;129(2):268-9.
62. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;353:i3140.
63. Marrugat J, Subirana I, Comin E, Cabezas C, Vila J, Elosua R, et al. Validity of an adaptation of the Framingham cardiovascular risk function: the VERIFICA Study. *J Epidemiol Community Health* 2007;61(1):40-7.
64. Andersson C, Enserro D, Larson MG, Xanthakis V, Vasan RS. Implications of the US cholesterol guidelines on eligibility for statin therapy in the community: comparison of observed and predicted risks in the Framingham Heart Study Offspring Cohort. *Journal of the American Heart Association* 2015;4(4).
65. Vergouwe Y, Nieboer D, Oostenbrink R, Debray TP, Murray GD, Kattan MW, et al. A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med* 2016.
66. Grundy SM, D'Agostino Sr RB, Mosca L, Burke GL, Wilson PW, Rader DJ, et al. Cardiovascular risk assessment based on US cohort studies: findings from a National Heart, Lung, and Blood institute workshop. *Circulation* 2001;104(4):491-6.
67. Hajifathalian K, Ueda P, Lu Y, Woodward M, Ahmadvand A, Aguilar-Salinas CA, et al. A novel risk score to predict cardiovascular disease risk in national populations (Globorisk): a pooled analysis of prospective cohorts and health examination surveys. *Lancet Diabetes Endocrinol* 2015;3(5):339-55.
68. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J* 2003;24(11):987-1003.
69. van Dis I, Kromhout D, Geleijnse JM, Boer JM, Verschuren WM. Evaluation of cardiovascular risk predicted by different SCORE equations: the Netherlands as an example. *Eur J Cardiovasc Prev Rehabil* 2010;17(2):244-9.
70. De Bacquer D, De Backer G. Predictive ability of the SCORE Belgium risk chart for cardiovascular mortality. *Int J Cardiol* 2010;143(3):385-90.

71. Sans S, Fitzgerald AP, Royo D, Conroy R, Graham I. [Calibrating the SCORE cardiovascular risk chart for use in Spain]. *Rev Esp Cardiol* 2007;60(5):476-85.
72. Usher-Smith JA, Silarova B, Schuit E, Gm Moons K, Griffin SJ. Impact of provision of cardiovascular disease risk estimates to healthcare professionals and patients: a systematic review. *BMJ Open* 2015;5(10):e008717.
73. Lackland DT, Elkind MS, D'Agostino R, Sr., Dhamoon MS, Goff DC, Jr., Higashida RT, et al. Inclusion of stroke in cardiovascular risk prediction instruments: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2012;43(7):1998-2027.
74. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;357:j2099.

Supplemental material

1 Supplementary introduction

1.1 Review question and PICOTS components

Review question: "What is the predictive performance of the Framingham Wilson, ATP III and PCE models in men and women separately for predicting 10-year risk of coronary heart disease (CHD) or cardiovascular disease (CVD) in the general population?"

Patients - General population, divided by gender. Include population based and primary care cohorts; exclude cohorts in which specific patient populations were excluded

Intervention and Comparators - Framingham Wilson 1998, Framingham ATP III 2003, PCE 2013, for men and women separately

Outcome - Outcome for which the original models were developed (fatal or nonfatal CHD for ATP III and Wilson, fatal or nonfatal CVD for PCE)

Timing/prediction horizon - 10 years

Setting - Primary care and public health

1.2 Overview of Framingham prediction models and PCE

	Framingham Wilson[1]	Framingham ATP III[2,3]	PCE[4]
Development cohort(s)	- Framingham Heart Study: 11th examination of the original Framingham cohort or initial examination of the Framingham Offspring Study	- Framingham Heart Study	- Framingham Heart Study: original and offspring cohorts. - Atherosclerosis Risk in Communities (ARIC) study - Cardiovascular Health Study (CHS) - Coronary Artery Risk Development in Young Adults (CARDIA) study
In/exclusion criteria	People aged 30 to 74 years old at the time of their Framingham Heart Study examination in 1971 to 1974. Persons with overt CHD at the baseline examination were excluded.	People aged 20 to 79 without diabetes.	People aged 40 to 79, apparently healthy, African American or White, and free of a previous history of myocardial infarction (recognized or unrecognized), stroke, congestive heart failure, percutaneous coronary intervention, coronary bypass surgery, or atrial fibrillation.

	Framingham Wilson[1]	Framingham ATP III[2,3]	PCE[4]
Predictors	Age Smoking Diabetes Systolic blood pressure Diastolic blood pressure Total or LDL cholesterol HDL cholesterol	Age Smoking Systolic blood pressure Treatment of blood pressure Total cholesterol HDL cholesterol	Age Smoking Diabetes Systolic blood pressure Treatment of blood pressure Total cholesterol HDL cholesterol
Predicted outcome	Fatal or nonfatal CHD, defined as angina pectoris, recognized and unrecognized myocardial infarction, coronary insufficiency, and coronary heart disease death.	Fatal or nonfatal CHD, defined as myocardial infarction or CHD death.	Atherosclerotic CVD defined as nonfatal myocardial infarction or coronary heart disease death, or fatal or nonfatal stroke.
Prediction horizon	10 years	10 years	10 years

2 Supplementary methods

2.1 Search strategy

2.1.1 MEDLINE search strategy

1 chd risk assessment$.mp.
2 cvd risk assessment$.mp.
3 heart disease risk assessment$.mp.
4 coronary disease risk assessment$.mp.
5 cardiovascular disease risk assessment$.mp.
6 cardiovascular risk assessment$.mp.
7 cv risk assessment$.mp.
8 cardiovascular disease$ risk assessment$.mp.
9 coronary risk assessment$.mp.
10 coronary risk scor$.mp.
11 heart disease risk scor$.mp.
12 chd risk scor$.mp.
13 cardiovascular risk scor$.mp.
14 cardiovascular disease$ risk scor$.mp.
15 cvd risk scor$.mp.
16 cv risk scor$.mp.

17 or/1-16
18 cardiovascular diseases/
19 coronary disease/
20 cardiovascular disease$.mp.
21 heart disease$.mp.
22 coronary disease$.mp.
23 cardiovascular risk?.mp.
24 coronary risk?.mp.
25 exp hypertension/
26 exp hyperlipidemia/
27 or/18-26
28 risk function.mp.
29 Risk Assessment/mt
30 risk functions.mp.
31 risk equation$.mp.
32 risk chart?.mp.
33 (risk adj3 tool$).mp.
34 risk assessment function?.mp.
35 risk assessor.mp.
36 risk appraisal$.mp.
37 risk calculation$.mp.
38 risk calculator$.mp.
39 risk factor$ calculator$.mp.
40 risk factor$ calculation$.mp.
41 risk engine$.mp.
42 risk equation$.mp.
43 risk table$.mp.
44 risk threshold$.mp.
45 risk disc?.mp.
46 risk disk?.mp.
47 risk scoring method?.mp.
48 scoring scheme?.mp.
49 risk scoring system?.mp.
50 risk prediction?.mp.
51 predictive instrument?.mp.
52 project$ risk?.mp.
53 cdss.mp.
54 or/28-53
55 27 and 54
56 17 or 55

57 new zealand chart$.mp.
58 sheffield table$.mp.
59 procam.mp.
60 General Rule to Enable Atheroma Treatment.mp.
61 dundee guideline$.mp.
62 shaper scor$.mp.
63 (brhs adj3 score$).mp.
64 (brhs adj3 risk$).mp.
65 copenhagen risk.mp.
66 precard.mp.
67 (framingham adj1 (function or functions)).mp.
68 (framingham adj2 risk).mp.
69 framingham equation.mp.
70 framingham model$.mp.
71 (busselton adj2 risk$).mp.
72 (busselton adj2 score$).mp.
73 erica risk score$.mp.
74 framingham scor$.mp.
75 dundee scor$.mp.
76 brhs scor$.mp.
77 British Regional Heart study risk scor$.mp.
78 brhs risk scor$.mp.
79 dundee risk scor$.mp.
80 framingham guideline$.mp.
81 framingham risk?.mp.
82 new zealand table$.mp.
83 ncep guideline?.mp.
84 smac guideline?.mp.
85 copenhagen risk?.mp.
86 or/57-85
87 56 or 86
88 exp decision support techniques/
89 Diagnosis, Computer-Assisted/
90 Decision Support Systems,Clinical/
91 algorithms/
92 algorithm?.mp.
93 algorythm?.mp.
94 decision support?.mp.
95 predictive model?.mp.
96 treatment decision?.mp.

97 scoring method$.mp.
98 (prediction$ adj3 method$).mp.
99 or/88-98
100 Risk Factors/
101 exp Risk Assessment/
102 (risk? adj1 assess$).mp.
103 risk factor?.mp.
104 or/100-103
105 27 and 99 and 104
106 87 or 105
107 stroke.mp.
108 exp Stroke/
109 cerebrovascular.mp. or exp Cerebrovascular Circulation/
110 limit 106 to ed=20040101-20130601
111 107 or 108 or 109
112 111 and 54
113 111 and 99 and 104
114 112 or 113
115 106 or 114

2.1.2 Citation search

Web of Science and Scopus were searched for studies citing the following references:
Wilson:
- Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97(18):1837-47.
ATP III:
- Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* 2002;106(25):3143-421.
- Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). *JAMA* 2001;285(19):2486-97.
PCE:
- Goff DC, Jr., Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014;129(25 Suppl 2):S49-73.
- Goff DC, Jr., Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/

American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol* 2014;63(25 Pt B):2935-59

2.2 Items for data extraction

List of items for which data were extracted.

Item	Description / examples
Validated model	Framingham Wilson, Framingham ATPIII, PCE; men or women; race (PCE); LDL or total cholesterol (Framingham Wilson).
Study type	Only external validation; external validation and development of a new model; external validation and incremental value assessment.
Study design	Cohort, randomized controlled trial
Eligibility criteria for participants	Age, (exclusion of) comorbidities, treatment, race.
Study dates	Inclusion dates, end of follow-up, follow-up time.
Prediction horizon	Time period for which predictions were made, e.g. 10 years.
Geographical location	Country and continent.
Case-mix	Information on the frequency, or mean/median and spread of the following population characteristics of the validation study: age, smoking, diabetes, treatment, hypertension, systolic blood pressure, diastolic blood pressure, total cholesterol, LDL cholesterol, HDL cholesterol, race, other diseases, linear predictor, 10-year predicted survival probability.
Predictors	Full definition, measurement method, blinding of measurements.
Predicted outcome	Full definition, including ICD-codes.
Sample size	Number of participants, number of events, Kaplan-Meier 10-year survival probability.
Performance	C-statistic, 10-year total observed/expected ratio, standard error, 95% confidence intervals, calibration plot, calibration table. Performance of the original model and after updating the model were extracted.

2.3 Formulas used to estimate missing quantitative information

Case-mix variables

For the case-mix variables age, systolic blood pressure (SBP), HDL cholesterol and total cholesterol, we needed the mean and standard deviation (sd) for our analyses, however some studies only reported the median and 25th and 75th percentiles, or the minimum and maximum. If the median and percentiles were reported, we used equation 14 from a paper by Wan et al. to approximate the mean, and equation 16 to approximate the sd.[5] If only the range was reported, we used equation 5 from the same paper to approximate the sd. One study reported the number of participants in SBP, HDL cholesterol and total cholesterol categories.[6] To estimate the mean and sd, we took bootstrap samples from a uniform distribution per category, with sample size equal to the number of participants in the original categories, and calculated the mean and sd of this sample. This process was repeated 1000 times, and subsequently the overall (average) mean and sd were calculated.

C-statistics

If the precision of the c-statistic was not reported, we estimated this from the c-statistic and sample size of the study, using the formula described by Newcombe and Hanley.[7,8]

OE ratio

Various equations were used to estimate the standard error of the OE ratio, depending on which information was reported. All equations (as numbered) are described in the appendix of Debray et al.[9] If the SE of the OE ratio was reported, we used equation 16 to estimate the SE of ln(OE), if the observed event risk (Po), the expected event risk (Pe), and the SE of Po were reported, we used equation 51, and if only Po and Pe were reported we used equation 27.

If the OE ratio was reported for a prediction horizon shorter than 10 years, we extrapolated Po and Pe separately to 10 years using the following equation based on the Poisson distribution:

$$S_{KM,10} = \exp\left(\frac{10 \ln(S_{KM,l})}{l}\right)$$

where $S_{KM,10}$ is the Kaplan Meier estimate of survival at 10 years, and $S_{KM,l}$ the Kaplan Meier survival estimate at time *l*. Po can be calculated by taking 1 - $S_{KM,10}$.

2.4 Statistical analyses

2.4.1 Meta-analysis

The logit c-statistic and log OE ratio were pooled using random-effects meta-analyses accounting for the presence of between-study heterogeneity, weighted by the inverse of the variance. The Hartung-Knapp-Sidik-Jonkman (HKSJ) method was used when calculating 95% confidence intervals.[10] The 95% prediction interval was calculated using the equation described by Debray et al.[9]

2.4.2 Calibration slope

The calibration slope can be calculated as follows:

$$O_{ij} \sim (N_{ij}, p_{ij})$$
$$logit(P_{ij}) = \alpha_i + \beta_i logit(P_{E,ij})$$
$$\beta_i \sim N(\mu_{cal.slope}, \tau^2_{cal.slope})$$

Where O_{ij} is the number of observed events in subgroup *j* of study *i*, modeled using a binomial distribution with event probability p_{ij}. The calibration slope is given by $\hat{\mu}_{cal.slope}$.

2.4.3 Meta-regression

To investigate if the performance of the six models was influenced by differences in, for example, study populations, we fitted meta-regression models with a single covariate. The following categorical covariates were considered:

- age range of included participants: comparable (if both the upper and lower limit were within 5 years of the age range in the development population), narrower (if the lower limit was more than 5 years higher and/or the upper limit was more than 5 years lower), younger (if the lower limit was more than 5 years lower), older (if the upper limit was more than 5 years higher) or not reported (NR),
- in- or exclusion of participants with diabetes at baseline,
- in- or exclusion of participants with CHD or CVD at baseline,
- continent,
- prediction horizon: <10 year, 10 year, >10 year or NR,
- type of model used: for Wilson LDL or total cholesterol, for PCE white and others, or African American.

The following continuous covariates were included: mean and standard deviation of age, systolic blood pressure, HDL and total cholesterol, year in which the recruitment of participants for the study started, and the prediction horizon.

2.4.4 Sensitivity analyses

We performed several sensitivity analyses. Firstly, we excluded all external validations with high risk of bias for at least one domain. Secondly, since almost all validations scored high risk of bias for either the domain sample size and participant flow or analysis, we performed a second analysis in which we only excluded external validations with high risk of bias for any of the three domains: participant selection, predictors, or outcome. Thirdly, we used the number of events rather than the inverse of the variance as weighting factor in the meta-analysis, as suggested by Pennells et al. to increase statistical power.[11] Fourthly, we fitted a bivariate model with both the c-statistic and the 10-year total OE ratio as outcomes.[12] Fifthly, we repeated the analyses with the original OE ratio without extrapolating it to 10 years.

3 Supplementary results

3.1 Description of excluded outcomes

The table below gives an overview of the validations that were excluded because the outcome definition differed too much from the definition used in model development.

Model	Reference	Outcome category	Outcome definition
Wilson men	Lee 2008[13]	Fatal CVD	All deaths due to ischaemic heart disease (ICD-9 410-414) and cerebrovascular accidents (ICD-9 430-438).
	Stork 2006[14]	Fatal CVD	Not reported
	Barroso 2010[15]	Fatal or nonfatal CVD	Angina and myocardial infarction (fatal and non-fatal), and fatal cardiovascular disease (cardiac death of coronary and non-coronary origin, death of cerebrovascular origin, and deaths from other cardiovascular causes).
Wilson women	Lee 2008[13]	Fatal CVD	All deaths due to ischaemic heart disease (ICD-9 410-414) and cerebrovascular accidents (ICD-9 430-438).
	Barroso 2010[15]	Fatal or nonfatal CVD	Angina and myocardial infarction (fatal and non-fatal), and fatal cardiovascular disease (cardiac death of coronary and non-coronary origin, death of cerebrovascular origin, and deaths from other cardiovascular causes).
	Ridker 2007[16]	Fatal or nonfatal CVD	Myocardial infarction, ischemic stroke, coronary revascularization, and cardiovascular deaths

Model	Reference	Outcome category	Outcome definition
	Ridker 2007[16]	Fatal or nonfatal CVD	Myocardial infarction, ischemic stroke, coronary revascularization, and cardiovascular deaths
ATP III men	Berry 2007[17]	Fatal CHD	Coronary heart disease mortality
	Berry 2007[17]	Fatal CHD	Coronary heart disease mortality
	Berry 2007[17]	Fatal CHD	Coronary heart disease mortality
	Berry 2007[17]	Fatal CHD	Coronary heart disease mortality
	Dunder 2004[18]	Fatal or nonfatal MI	Hospitalization or death due to myocardial infarction (ICD 410/I 21).
	Ridker 2007[16]	Fatal or nonfatal CVD	Myocardial infarction, ischemic stroke, coronary revascularization, and cardiovascular deaths

CVD: Cardiovascular disease, ICD: International Classification of Diseases, CHD: coronary heart disease, ATP: Adult treatment panel, MI: myocardial infarction

3.2 Cohorts used multiple times to validate the same model

Below an overview is given of the cohorts that were used more than once to validate the same model, with rationale for the choice of cohort that was kept in the analyses, separately for validations included in the meta-analyses of calibration and discrimination.

OE ratio:

Reference	Cohort	Model	Decision	Explanation
Jung 2015[19]	Korean Heart Study	PCE men African American	Excluded	AHA guidelines advice to use the white model for this group of people
Jung 2015[19]		PCE men white	Included	
Jung 2015[19]	Korean Heart Study	PCE women African American	Excluded	AHA guidelines advice to use the white model for this group of people
Jung 2015[19]		PCE women white	Included	

Reference	Cohort	Model	Decision	Explanation
De Filippis 2015[20]	MESA study	PCE men	Excluded	Most general population, fits review question best, most up-to-date population
De Filippis 2017[21]		PCE men	Included	
Goff 2014[4]		PCE men African American	Excluded	
Goff 2014[4]		PCE men white	Excluded	
De Filippis 2015[20]	MESA study	PCE women	Excluded	Most general population, fits review question best, most up-to-date population
De Filippis 2017[21]		PCE women	Included	
Goff 2014[4]		PCE women African American	Excluded	
Goff 2014[4]		PCE women white	Excluded	
Muntner 2014[22]	REGARDS study	PCE men	Included	Most general population, fits review question best
Goff 2014[4]		PCE men African American	Excluded	
Goff 2014[4]		PCE men white	Excluded	
Muntner 2014[22]	REGARDS study	PCE women	Included	Most general population, fits review question best
Goff 2014[4]		PCE women African American	Excluded	
Goff 2014[4]		PCE women white	Excluded	
Yang 2016[23]	China MUCA (1992)	PCE men African American	Excluded	AHA guidelines advice to use the white model for this group of people
Yang 2016[23]		PCE men white	Included	
Yang 2016[23]	China MUCA (1992)	PCE women African American	Excluded	AHA guidelines advice to use the white model for this group of people
Yang 2016[23]		PCE women white	Included	
Yang 2016[23]	CIMIC	PCE men African American	Excluded	AHA guidelines advice to use the white model for this group of people
Yang 2016[23]		PCE men white	Included	
Yang 2016[23]	CIMIC	PCE women African American	Excluded	AHA guidelines advice to use the white model for this group of people
Yang 2016[23]		PCE women white	Included	

Reference	Cohort	Model	Decision	Explanation
Yang 2016[23]	InterASIA and China MUCA (1998)	PCE men African American	Excluded	AHA guidelines advice to use the white model for this group of people
Yang 2016[23]		PCE men white	Included	
Yang 2016[23]	InterASIA and China MUCA (1998)	PCE women African American	Excluded	AHA guidelines advice to use the white model for this group of people
Yang 2016[23]		PCE women white	Included	
Mortensen 2015[24]	Copenhagen General Population Study	PCE men	Excluded	Most recent data
Mortensen 2017[25]		PCE men	Included	
Mortensen 2015[24]	Copenhagen General Population Study	PCE women	Excluded	Most recent data
Mortensen 2017[25]		PCE women	Included	

C-statistic:

Reference	Cohort	Model	Excluded	Explanation for decision
Mainous 2007[6]	ARIC study	Wilson men Total cholesterol	Included	Most general population, fits review question best
D'Agostino 2001[26]		Wilson men Total cholesterol	Excluded	
D'Agostino 2001[26]		Wilson men Total cholesterol	Excluded	
Mainous 2007[6]	ARIC study	Wilson women Total cholesterol	Included	Most general population, fits review question best
D'Agostino 2001[26]		Wilson women Total cholesterol	Excluded	
D'Agostino 2001[26]		Wilson women Total cholesterol	Excluded	
Jung 2015[19]	Korean Heart Study	PCE men African American	Excluded	AHA guidelines advice to use the white model for this group of people
Jung 2015[19]		PCE men white	Included	
Jung 2015[19]	Korean Heart Study	PCE women African American	Excluded	AHA guidelines advice to use the white model for this group of people
Jung 2015[19]		PCE women white	Included	
DeFilippis 2015[20]	MESA study	PCE men	Excluded	Most general population, fits review question best, most up-to-date population
De Filippis 2017[21]		PCE men	Included	
Goff 2014[4]		PCE men African American	Excluded	
Goff 2014[4]		PCE men white	Excluded	
DeFilippis 2015[20]	MESA study	PCE women	Excluded	Most general population, fits review question best, most up-to-date population
De Filippis 2017[21]		PCE women	Included	
Goff 2014[4]		PCE women African American	Excluded	
Goff 2014[4]		PCE women white	Excluded	
Muntner 2014[22]	REGARDS study	PCE men	Included	Most general population, fits review question best
Goff 2014[4]		PCE men African American	Excluded	
Goff 2014[4]		PCE men white	Excluded	

Reference	Cohort	Model	Excluded	Explanation for decision
Muntner 2014[22]	REGARDS study	PCE women	Included	Most general population, fits review question best
Goff 2014[4]		PCE women African American	Excluded	
Goff 2014[4]		PCE women white	Excluded	
Koller 2012[27]	Rotterdam Study	ATP III men	Included	Most recent publication
Koller 2007[28]		ATP III men	Excluded	
Koller 2012[27]	Rotterdam Study	ATP III women	Included	Most recent publication
Koller 2007[28]		ATP III women	Excluded	
Yang 2016[23]	China MUCA (1992)	PCE men African American	Excluded	AHA guidelines advice to use the white model for this group of people
Yang 2016[23]		PCE men white	Included	
Yang 2016[23]	China MUCA (1992)	PCE women African American	Excluded	AHA guidelines advice to use the white model for this group of people
Yang 2016[23]		PCE women white	Included	
Yang 2016[23]	CIMIC	PCE men African American	Excluded	AHA guidelines advice to use the white model for this group of people
Yang 2016[23]		PCE men white	Included	
Yang 2016[23]	CIMIC	PCE women African American	Excluded	AHA guidelines advice to use the white model for this group of people
Yang 2016[23]		PCE women white	Included	
Yang 2016[23]	InterASIA and China MUCA (1998)	PCE men African American	Excluded	AHA guidelines advice to use the white model for this group of people
Yang 2016[23]		PCE men white	Included	
Yang 2016[23]	InterASIA and China MUCA (1998)	PCE women African American	Excluded	AHA guidelines advice to use the white model for this group of people
Yang 2016[23]		PCE women white	Included	
Mortensen 2015[24]	Copenhagen General Population Study	PCE men	Excluded	Most recent data
Mortensen 2017[25]		PCE men	Included	
Mortensen 2015[24]	Copenhagen General Population Study	PCE women	Excluded	Most recent data
Mortensen 2017[25]		PCE women	Included	

3.3 Characteristics of included validations

Table S1: characteristics of included external validations

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
Andersson 2015[29]	PCE men white	1971-1998	10/10	Framingham Heart Study Offspring Cohort	Framingham	Fatal or nonfatal CVD	284/3396	53.3 (40-75)	0.720 (0.014)	0.840 (0.050)
Andersson 2015[29]	PCE women white	1971-1998	10/10	Framingham Heart Study Offspring Cohort	Framingham	Fatal or nonfatal CVD	112/3838	53.1 (40-75)	0.773 (0.023)	0.674 (0.070)
Buitrago 2011[30]	Wilson men Total cholesterol	1994-2004	NR/10	Patients ascribed to the La Paz healthcare centre in Badajoz, Spain	Spain	Fatal or nonfatal CHD	22/201	50.9 (35-74)	0.630 (0.061)	0.673 (0.136)
Buitrago 2011[30]	Wilson women Total cholesterol	1994-2004	NR/10	Patients ascribed to the La Paz healthcare centre in Badajoz, Spain	Spain	Fatal or nonfatal CHD	8/246	53.6 (35-74)	0.650 (0.110)	0.423 (0.146)
Chia 2014[31]	PCE men white	1998-1998	NR/10	Patients registered with an outpatient primary care clinic of University Malaya Medical Centre	Malaysia	Fatal or nonfatal CVD	22/307	58.7 (40-79)	0.550 (0.050)	0.341 (0.070)

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
Chia 2014[31]	PCE women white	1998-1998	NR/10	Patients registered with an outpatient primary care clinic of University Malaya Medical Centre	Malaysia	Fatal or nonfatal CVD	23/615	56.9 (40-79)	0.610 (0.060)	0.552 (0.114)
Comin 2007[32]	Wilson men Total cholesterol	1995-1998	5/5	Patients from 67 health centers in autonomous Spanish regions	Spain	Fatal or nonfatal CHD	137/3285	55.7 (35-74)	0.679 (0.023)	0.387 (0.038)**
Comin 2007[32]	Wilson women Total cholesterol	1995-1998	5/5	Patients from 67 health centers in autonomous Spanish regions	Spain	Fatal or nonfatal CHD	86/3285	56.8 (35-74)	0.729 (0.030)	0.359 (0.048)**
Cook 2014[33]	PCE women	1992-1995	10.2/10	Womens Health Study	United States	Fatal or nonfatal CVD	632/27542	54.2 (45-79)	NR	0.611 (0.025)
Cooper 2005[34]	ATP III men	1989-NR	10.8/10	Second Northwick Park Heart Study	United Kingdom	Fatal or nonfatal CHD	219/2732	NR (50-64)	0.620 (0.020)	0.470 (NR)
D'Agostino 2001[26]†	Wilson men Total cholesterol	1987-1988	NR/5	ARIC study	United States	Fatal or nonfatal CHD	149/4705	54.6 (44-66)	0.750 (0.020)	0.931 (0.074)*

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
D'Agostino 2001[26]†	Wilson men Total cholesterol	1987-1988	NR/5	ARIC study	United States	Fatal or nonfatal CHD	46/1428	53.7 (44-66)	0.670 (0.040)	0.895 (0.127)*
D'Agostino 2001[26]	Wilson men Total cholesterol	1982-1982	NR/5	Physicians Health Study	United States	Fatal or nonfatal CHD	182/901	57.6 (40-74)	0.630 (0.023)	NR
D'Agostino 2001[26]	Wilson men Total cholesterol	1980-1982	NR/5	Honolulu Heart Program	United States	Fatal or nonfatal CHD	77/2755	61.9 (51-81)	0.720 (0.029)	0.466 (0.051)*
D'Agostino 2001[26]	Wilson men Total cholesterol	1965-1968	NR/5	Puerto Rico Heart Health Program	Puerto Rico	Fatal or nonfatal CHD	107/8713	54.1 (35-74)	0.690 (0.026)	0.352 (0.033)*
D'Agostino 2001[26]	Wilson men Total cholesterol	1989-1991	NR/5	Strong Heart Study	United States	Fatal or nonfatal CHD	46/1527	55.4 (45-75)	0.690 (0.039)	0.698 (0.097)*
D'Agostino 2001[26]	Wilson men Total cholesterol	1989-1990	NR/5	Cardiovascular Health Study	United States	Fatal or nonfatal CHD	71/956	69.7 (65-74)	0.630 (0.034)	NR
D'Agostino 2001[26]†	Wilson women Total cholesterol	1987-1988	NR/5	ARIC study	United States	Fatal or nonfatal CHD	52/5712	53.9 (44-66)	0.830 (0.029)	0.816 (0.11)*

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
D'Agostino 2001[26]†	Wilson women Total cholesterol	1987-1988	NR/5	ARIC study	United States	Fatal or nonfatal CHD	38/2333	53.3 (44-66)	0.790 (0.037)	1.069 (0.163)*
D'Agostino 2001[26]	Wilson women Total cholesterol	1989-1991	NR/5	Strong Heart Study	United States	Fatal or nonfatal CHD	23/2255	56.5 (45-75)	0.750 (0.051)	0.425 (0.082)*
D'Agostino 2001[26]	Wilson women Total cholesterol	1989-1990	NR/5	Cardiovascular Health Study	United States	Fatal or nonfatal CHD	44/1601	69.3 (65-74)	0.660 (0.041)	NR
De Filippis 2015[20]	Wilson men Total cholesterol	2000-2002	NR/10	MESA study	United States	Fatal or nonfatal CHD	164/1961	61.5 (50-74)	0.690 (0.020)	0.640 (0.046)
De Filippis 2015[20]	Wilson women Total cholesterol	2000-2002	NR/10	MESA study	United States	Fatal or nonfatal CHD	99/2266	61.5 (50-74)	0.600 (0.028)	0.680 (0.064)
De Filippis 2015[20]	ATP III men	2000-2002	NR/10	MESA study	United States	Fatal or nonfatal CHD	86/1961	61.5 (50-74)	0.710 (0.027)	0.386 (0.04)
De Filippis 2015[20]	ATP III wo men	2000-2002	NR/10	MESA study	United States	Fatal or nonfatal CHD	48/2266	61.5 (50-74)	0.670 (0.039)	0.693 (0.094)

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
De Filippis 2015[20]	PCE men	2000-2002	NR/10	MESA study	United States	Fatal or nonfatal CVD	125/1961	61.5 (50-74)	0.710 (0.021)	0.531 (0.044)
De Filippis 2015[20]	PCE women	2000-2002	NR/10	MESA study	United States	Fatal or nonfatal CVD	93/2266	61.5 (50-74)	0.700 (0.027)	0.599 (0.058)
De Filippis 2017[21]	PCE men	2000-2002	NR/10	MESA study	United States	Fatal or nonfatal CVD	220/3053	NR (45-79)	0.710 (0.018)	0.520 (0.034)
De Filippis 2017[21]	PCE women	2000-2002	NR/10	MESA study	United States	Fatal or nonfatal CVD	149/3388	NR (45-79)	0.740 (0.021)	0.500 (0.040)
De Las Heras Gala 2016[35]	PCE men white	1994-2001	NR/10	Kora study	Germany	Fatal or nonfatal CVD	257/2584	56.4 (40-79)	0.736 (0.015)	0.700 (0.042)
De Las Heras Gala 2016[35]	PCE women white	1994-2001	NR/10	Kora study	Germany	Fatal or nonfatal CVD	126/2654	55.5 (40-79)	0.809 (0.019)	0.800 (0.070)
De Las Heras Gala 2016[35]	PCE men white	2000-2003	NR/10	HNR study	Germany	Fatal or nonfatal CVD	186/2005	58.8 (40-79)	0.670 (0.020)	0.620 (0.043)

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
De Las Heras Gala 2016[35]	PCE women white	2000-2003	NR/10	HNR study	Germany	Fatal or nonfatal CVD	84/2203	59.1 (40-79)	0.756 (0.026)	0.590 (0.062)
Emdin 2017[36]	PCE men	2008-2009	2.7/10	BioImage study	United States	Fatal or nonfatal CVD	43/1635	NR (55-80)	0.630 (0.044)	0.410 (0.062)
Emdin 2017[36]	PCE women	2008-2009	2.7/10	BioImage study	United States	Fatal or nonfatal CVD	31/2000	NR (60-80)	0.630 (0.031)	0.330 (0.067)
Empana 2003[37]	Wilson men LDL cholesterol	1991-1993	NR/5	PRIME study	Northern Ireland	Fatal or nonfatal CHD	120/2399	NR (50-59)	0.660 (0.025)	0.761 (0.069)*
Empana 2003[37]	Wilson men LDL cholesterol	1991-1993	NR/5	PRIME study	France	Fatal or nonfatal CHD	197/7359	NR (50-59)	0.680 (0.019)	0.422 (0.030)*
Ferrario 2005[38]	Wilson men Total cholesterol	1983-1996	9.1/10	CUORE study	Italy	Fatal or nonfatal CHD	312/6865	50.8 (35-69)	0.723 (0.028)	0.374 (0.019)
Goff 2014[4]	PCE men white	NR	NR/10	ARIC study, Framingham Heart Study	Framingham	Fatal or nonfatal CVD	539/5041	NR (40-79)	0.684 (0.012)	0.727 (0.028)

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
Goff 2014[4]†‡	PCE men white	2000-2002	NR/6	MESA study	United States	Fatal or nonfatal CVD	57/1184	NR (40-79)	0.704 (0.035)	0.636 (0.080)*
Goff 2014[4]†‡	PCE men white	2003-2007	NR/4	REGARDS study	United States	Fatal or nonfatal CVD	218/5296	NR (40-79)	0.595 (0.020)	0.823 (0.051)*
Goff 2014[4]	PCE men African American	NR	NR/10	ARIC study, Framingham Heart Study	Framingham	Fatal or nonfatal CVD	107/735	NR (40-79)	0.711 (0.027)	0.944 (0.081)
Goff 2014[4]†‡	PCE men African American	2000-2002	NR/6	MESA study	United States	Fatal or nonfatal CVD	36/799	NR (40-79)	0.669 (0.046)	0.538 (0.085)*
Goff 2014[4]†‡	PCE men African American	2003-2007	NR/4	REGARDS study	United States	Fatal or nonfatal CVD	136/2969	NR (40-79)	0.556 (0.025)	0.904 (0.069)*
Goff 2014[4]	PCE women white	NR	NR/10	ARIC study, Framingham Heart Study	Framingham	Fatal or nonfatal CVD	400/6509	NR (40-79)	0.738 (0.013)	0.777 (0.036)
Goff 2014[4]†‡	PCE women white	2000-2002	NR/6	MESA study	United States	Fatal or nonfatal CVD	37/1273	NR (40-79)	0.711 (0.043)	0.772 (0.123)*

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
Goff 2014[4]†‡	PCE women white	2003-2007	NR/4	REGARDS study	United States	Fatal or nonfatal CVD	101/6333	NR (40-79)	0.660 (0.027)	0.787 (0.071)*
Goff 2014[4]	PCE women African American	NR	NR/10	ARIC study, Framingham Heart Study	Framingham	Fatal or nonfatal CVD	127/1367	NR (40-79)	0.707 (0.024)	0.944 (0.078)
Goff 2014[4]†‡	PCE women African American	2000-2002	NR/6	MESA study	United States	Fatal or nonfatal CVD	28/978	NR (40-79)	0.768 (0.045)	0.512 (0.092)*
Goff 2014[4]†‡	PCE women African American	2003-2007	NR/4	REGARDS study	United States	Fatal or nonfatal CVD	126/5275	NR (40-79)	0.662 (0.024)	0.683 (0.056)*
Jee 2014[39]	Wilson men Total cholesterol	1996-2001	11.6/10	Korean Heart Study	South Korea	Fatal or nonfatal CHD	2086/164005	45.8 (30-74)	NR	NR
Jee 2014[39]	Wilson women Total cholesterol	1996-2001	11.6/10	Korean Heart Study	South Korea	Fatal or nonfatal CHD	510/104310	47.6 (30-74)	NR	NR
Jung 2015[19]	PCE men wh ite	1996-2001	NR/10	Korean Heart Study	South Korea	Fatal or nonfatal CVD	7669/114622	50.1 (40-79)	0.727 (0.003)	0.634 (0.008)

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
Jung 2015[19]†‡	PCE men African American	1996-2001	NR/10	Korean Heart Study	South Korea	Fatal or nonfatal CVD	7669/114622	50.1 (40-79)	0.725 (0.003)	1.346 (0.023)
Jung 2015[19]	PCE women white	1996-2001	NR/10	Korean Heart Study	South Korea	Fatal or nonfatal CVD	4658/77983	51.8 (40-79)	0.738 (0.004)	0.570 (0.007)
Jung 2015[19]†‡	PCE women African American	1996-2001	NR/10	Korean Heart Study	South Korea	Fatal or nonfatal CVD	4658/77983	51.8 (40-79)	0.739 (0.004)	0.754 (0.013)
Kavousi 2014[40]	ATP III men	1997-2001	NR/10	Rotterdam Study	Netherlands	Fatal or nonfatal CHD	98/1431	64.9 (55-75)	0.670 (0.026)	0.422 (0.043)
Kavousi 2014[40]	ATP III wo men	1997-2001	NR/10	Rotterdam Study	Netherlands	Fatal or nonfatal CHD	62/1976	65.1 (55-75)	0.690 (0.031)	0.574 (0.076)
Kavousi 2014[40]	PCE men	1997-2001	NR/10	Rotterdam Study	Netherlands	Fatal or nonfatal CVD	192/1513	65.1 (55-75)	0.670 (0.02)	0.591 (0.04)
Kavousi 2014[40]	PCE women	1997-2001	NR/10	Rotterdam Study	Netherlands	Fatal or nonfatal CVD	151/1920	65.2 (55-75)	0.680 (0.023)	0.681 (0.055)

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
Khalili 2015[41]	PCE men white	1999-2001	10.1/10	Tehran Lipid and Glucose Study (TLGS)	Iran	Fatal or nonfatal CVD	200/2353	54.6 (40-75)	0.740 (0.018)	0.758 (0.053)
Khalili 2015[41]	PCE women white	1999-2001	10.1/10	Tehran Lipid and Glucose Study (TLGS)	Iran	Fatal or nonfatal CVD	98/2749	52.5 (40-75)	0.820 (0.021)	0.839 (0.086)
Koller 2007[28]†	ATP III men	1990-1993	12.9/10	Rotterdam Study	Netherlands	Fatal or nonfatal CHD	351/2452	68.5 (55-NR)	0.630 (0.057)	0.722 (0.039)
Koller 2007[28]†	ATP III women	1990-1993	12.9/10	Rotterdam Study	Netherlands	Fatal or nonfatal CHD	448/4343	71.1 (55-NR)	0.730 (0.049)	0.980 (0.048)
Koller 2012[27]	ATP III men	1990-1993	14.9/10	Rotterdam Study	Netherlands	Fatal or nonfatal CHD	283/1454	73.3 (65-NR)	0.600 (0.018)	NR
Koller 2012[27]	ATP III men	1989-1992	16.5/10	Cardiovascular Health Study	United States	Fatal or nonfatal CHD	563/1917	72.7 (65-NR)	0.600 (0.015)	NR
Koller 2012[27]	ATP III women	1990-1993	14.9/10	Rotterdam Study	Netherlands	Fatal or nonfatal CHD	415/2849	76.3 (65-NR)	0.650 (0.018)	NR

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
Koller 2012[27]	ATP III women	1989-1992	16.5/10	Cardiovascular Health Study	United States	Fatal or nonfatal CHD	603/3029	71.7 (65-NR)	0.660 (0.013)	NR
Lee 2015[42]	PCE men white	1995-2004	10/10	Hong Kong Cardiovascular Risk Factor Prevalence Study (CRISPS) cohort	China	Fatal or nonfatal CVD	80/679	55.8 (40-74)	0.714 (0.049)	1.054 (0.102)
Lee 2015[42]	PCE women white	1995-2004	10/10	Hong Kong Cardiovascular Risk Factor Prevalence Study (CRISPS) cohort	China	Fatal or nonfatal CVD	42/797	53.4 (40-74)	0.765 (0.039)	1.438 (0.191)
Lloyd-Jones 2004[43]	Wilson men Total cholesterol	1971-NR	NR/10	Framingham Heart Study	Framingham	Fatal or nonfatal CHD	NR/2716	NR (40-94)	NR	NR
Lloyd-Jones 2004[43]	Wilson women Total cholesterol	1971-NR	NR/10	Framingham Heart Study	Framingham	Fatal or nonfatal CHD	NR/3500	NR (40-94)	NR	NR
Mainous 2007[6]	Wilson men Total cholesterol	1987-1989	NR/10	ARIC study	United States	Fatal or nonfatal CHD	NR/6239	54.4 (45-64)	0.691 (0.011)	NR

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
Mainous 2007[6]	Wilson women Total cholesterol	1987-1989	NR/10	ARIC study	United States	Fatal or nonfatal CHD	NR/8104	53.8 (45-64)	0.808 (0.008)	NR
Marrugat 2007[44]	Wilson men Total cholesterol	1995-1998	NR/5	VERIFICA study	Spain	Fatal or nonfatal CHD	98/2447	55.7 (35-74)	0.680 (0.024)	0.407 (0.040)*
Marrugat 2007[44]	Wilson women Total cholesterol	1995-1998	NR/5	VERIFICA study	Spain	Fatal or nonfatal CHD	56/3285	56.8 (35-74)	0.730 (0.030)	0.395 (0.053)*
Mortensen 2015[24]	PCE men	2003-2008	NR/5	Copenhagen General Population Study	Denmark	Fatal or nonfatal CVD	467/16398	56 (40-75)	0.647 (0.013)	0.597 (0.027)*
Mortensen 2015[24]	PCE women	2003-2008	NR/5	Copenhagen General Population Study	Denmark	Fatal or nonfatal CVD	367/21494	55.7 (40-75)	0.669 (0.014)	1.058 (0.055)*
Mortensen 2017[25]	PCE men	2003-2009	NR/5	Copenhagen General Population Study	Denmark	Fatal or nonfatal CVD	1205/19383	56 (40-75)	0.710 (0.008)	0.661 (NR)
Mortensen 2017[25]	PCE women	2003-2009	NR/5	Copenhagen General Population Study	Denmark	Fatal or nonfatal CVD	1012/25506	56 (40-75)	0.710 (0.008)	1.280 (NR)

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
Muntner 2014[22]	PCE men	2003-2007	NR/5	REGARDS study	United States	Fatal or nonfatal CVD	376/NR	NR (45-79)	0.650 (0.015)	0.721 (0.035)
Muntner 2014[22]	PCE women	2003-2007	NR/5	REGARDS study	United States	Fatal or nonfatal CVD	298/NR	NR (45-79)	0.740 (0.013)	0.813 (0.044)
Pike 2016[45]	PCE men	2005-2012	NR/10	Mayo Clinic Biobank	United States	Fatal or nonfatal CVD	246/3093	59 (30-75)	0.630 (0.018)	0.610 (0.037)
Pike 2016[45]	PCE women	2005-2012	NR/10	Mayo Clinic Biobank	United States	Fatal or nonfatal CVD	247/5690	56 (30-75)	0.690 (0.015)	0.610 (0.038)
Rana 2016[46]	PCE men	2008-2008	NR/5	Kaiser Permanente Northern California	United States	Fatal or nonfatal CVD	NR/118080	NR (40-75)	0.680 (NR)	NR
Rana 2016[46]	PCE women	2008-2008	NR/5	Kaiser Permanente Northern California	United States	Fatal or nonfatal CVD	NR/189511	NR (40-75)	0.720 (NR)	NR
Reissigova 2007[47]	Wilson men Total cholesterol	1975-1979	NR/10	Primary Prevention Study of Atherosclerotic Risk Factors (STULONG)	Czech Republic	Fatal or nonfatal CHD	83/646	51.2 (38-49)	0.638 (0.027)	0.217 (0.019)

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
Rodondi 2012[48]	Wilson men Total cholesterol	1997-1998	8.3/7.5	Health ABC Study	United States	Fatal or nonfatal CHD	205/981	73.6 (70-79)	0.583 (0.024)	1.083 (0.068)*
Rodondi 2012[48]	Wilson women Total cholesterol	1997-1998	8.3/7.5	Health ABC Study	United States	Fatal or nonfatal CHD	146/1212	73.4 (70-79)	0.577 (0.028)	2.049 (0.167)*
Ryckman 2015[49]	Wilson men Unclear	2004-2005	NR/NR	Series of adults undergoing colorectal cancer screening	United States	Fatal or nonfatal CHD	NR/NR	NR (NR)	NR	NR
Ryckman 2015[49]	Wilson women Unclear	2004-2005	NR/NR	Series of adults undergoing colorectal cancer screening	United States	Fatal or nonfatal CHD	NR/NR	NR (NR)	NR	NR
Simmons 2008[50]	Wilson men Total cholesterol	1993-1998	NR/10	EPIC-Norfolk	United Kingdom	Fatal or nonfatal CHD	430/4513	58.3 (40-79)	0.710 (0.010)	0.546 (0.025)
Simmons 2008[50]	Wilson women Total cholesterol	1993-1998	NR/10	EPIC-Norfolk	United Kingdom	Fatal or nonfatal CHD	250/5782	57.6 (40-79)	0.710 (0.015)	0.560 (0.036)
Simons 2003[51]	ATP III men	1988-1989	NR/10	Dubbo Study	Australia	Fatal or nonfatal CHD	105/755	NR (60-79)	NR	0.954 (0.086)

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
Simons 2003[51]	ATP III wo men	1988-1989	NR/10	Dubbo Study	Australia	Fatal or nonfatal CHD	80/1045	NR (60-79)	NR	0.899 (0.096)
Suka 2001[52]	Wilson men Total cholesterol	1991-1993	NR/NR	Employee health management center in a Japanese company	Japan	Fatal or nonfatal CHD	80/5611	44.7 (30-59)	0.710 (0.029)	NR
Sussman 2017[53]	PCE men	2007-NR	NR/5	US Department of Veterans Affairs	United States	Fatal or nonfatal CVD	80412/ 1435937	62 (45-80)	0.657 (0.001)	0.627 (0.002)*
Sussman 2017[53]	PCE women	2007-NR	NR/5	US Department of Veterans Affairs	United States	Fatal or nonfatal CVD	1599/76155	55.6 (45-80)	0.726 (0.006)	0.914 (0.023)*
Vaidya 2007[54]	Wilson men Total cholesterol	1983-1996	NR/10	10 Baltimore area hospitals	United States	Fatal or nonfatal CHD	81/404	45.2 (30-59)	0.698 (0.03)	1.701 (0.170)
Vaidya 2007[54]	Wilson women Total cholesterol	1983-1996	NR/10	10 Baltimore area hospitals	United States	Fatal or nonfatal CHD	27/380	46.1 (30-59)	0.787 (0.04)	1.141 (0.212)
Yang2016[23]	PCE men wh ite	1998-2001	NR/10	InterASIA and China MUCA (1998)	China	Fatal or nonfatal CVD	451/10334	48.8 (35-74)	0.762 (0.011)	0.657 (0.030)

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
Yang2016[23]	PCE men white	1992-1994	NR/10	China MUCA (1992)	China	Fatal or nonfatal CVD	216/6565	46.5 (35-59)	0.768 (0.018)	0.649 (0.043)
Yang2016[23]	PCE men white	2007-2008	NR/5	CIMIC	China	Fatal or nonfatal CVD	755/26872	55.3 (16-99)	0.761 (0.009)	0.636 (0.023)*
Yang2016[23]	PCE women white	1998-2001	NR/10	InterASIA and China MUCA (1998)	China	Fatal or nonfatal CVD	285/10986	48.4 (35-74)	0.783 (0.014)	1.102 (0.064)
Yang2016[23]	PCE women white	1992-1994	NR/10	China MUCA (1992)	China	Fatal or nonfatal CVD	168/7558	46.6 (35-59)	0.786 (0.017)	1.368 (0.104)
Yang2016[23]	PCE women white	2007-2008	NR/5	CIMIC	China	Fatal or nonfatal CVD	738/43966	53.9 (16-99)	0.785 (0.007)	1.110 (0.041)*
Yang2016[23]	PCE men African American	1998-2001	NR/10	InterASIA and China MUCA (1998)	China	Fatal or nonfatal CVD	451/10334	48.8 (35-74)	0.769 (0.011)	0.562 (0.026)
Yang2016[23]	PCE men African American	1992-1994	NR/10	China MUCA (1992)	China	Fatal or nonfatal CVD	216/6565	46.5 (35-59)	0.790 (0.017)	0.482 (0.032)

Table S1: Continued

Reference	Validated model	Recruitment years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n participants	Mean age (range)	C (SE)	OE (SE)
Yang2016[23]	PCE men African American	2007-2008	NR/5	CIMIC	China	Fatal or nonfatal CVD	755/26872	55.3 (16-99)	0.750 (0.008)	0.600 (0.022)*
Yang2016[23]	PCE women African American	1998-2001	NR/10	InterASIA and China MUCA (1998)	China	Fatal or nonfatal CVD	285/10986	48.4 (35-74)	0.796 (0.013)	0.715 (0.042)
Yang2016[23]	PCE women African American	1992-1994	NR/10	China MUCA (1992)	China	Fatal or nonfatal CVD	168/7558	46.6 (35-59)	0.807 (0.016)	0.794 (0.061)
Yang2016[23]	PCE women African American	2007-2008	NR/5	CIMIC	China	Fatal or nonfatal CVD	738/43966	53.9 (16-99)	0.792 (0.007)	0.699 (0.026)*

* OE ratio extrapolated to 10 years
** OE ratio and corresponding SE extrapolated to 10 years
† Not included in analyses of c-statistic because model was validated more than once in the same cohort
‡ Not included in analyses of OE ratio because model was validated more than once in the same cohort
FU: follow-up, N: number, C: c-statistic, OE: observed:expected ratio, SE: standard error, NR: Not reported, CHD: coronary heart disease, CVD: cardiovascular disease.

Table S2: A summary of the reported case-mix in the included validation studies

	Wilson men	Wilson women	ATPIII men	ATPIII women	PCE men	PCE women
Total N	23	15	7	6	30	31
Eligibility age - comparable	6 (26.1%)	4 (26.7%)	0 (0.0%)	0 (0.0%)	22 (73.3%)	23 (74.2%)
Eligibility age - younger	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.3%)	1 (3.2%)
Eligibility age - older	2 (8.7%)	1 (6.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Eligibility age - narrower	14 (60.9%)	9 (60.0%)	4 (57.1%)	3 (50.0%)	5 (16.7%)	5 (16.1%)
Eligibility age - broader	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (6.7%)	2 (6.5%)
Eligibility age - NR	1 (4.3%)	1 (6.7%)	3 (42.9%)	3 (50.0%)	0 (0.0%)	0 (0.0%)
Eligibility CHD - not excluded	6 (26.1%)	4 (26.7%)	0 (0.0%)	0 (0.0%)	1 (3.3%)	1 (3.2%)
Eligibility CHD - CHD excl	9 (39.1%)	6 (40.0%)	0 (0.0%)	0 (0.0%)	1 (3.3%)	1 (3.2%)
Eligibility CHD - CVD excl	8 (34.8%)	5 (33.3%)	7 (100.0%)	6 (100.0%)	28 (93.3%)	29 (93.5%)
Eligibility diabetes - not excl	20 (87.0%)	13 (86.7%)	4 (57.1%)	3 (50.0%)	26 (86.7%)	27 (87.1%)
Eligibility diabetes - excl	3 (13.0%)	2 (13.3%)	3 (42.9%)	3 (50.0%)	4 (13.3%)	4 (12.9%)
Treated individuals - not excl	21 (91.3%)	14 (93.3%)	5 (71.4%)	5 (83.3%)	20 (66.7%)	22 (71.0%)

Table S2: Continued

	Wilson men	Wilson women	ATPIII men	ATPIII women	PCE men	PCE women
Treated individuals - excl	2 (8.7%)	1 (6.7%)	2 (28.6%)	1 (16.7%)	10 (33.3%)	9 (29.0%)
Age mean	58.0 (54.6-73.6), NR=4	57.6 (56.5-73.4), NR=2	72.7 (68.5-73.3), NR=2	71.7 (71.1-76.3), NR=1	58.7 (55.6-65.1), NR=10	56.0 (53.9-65.2), NR=10
Age sd	9.6 (7.4-13.5), NR=1	9.9 (7.4-13.5), NR=1	6.5 (5.6-7.1), NR=1	7.1 (5.9-9.6), NR=1	9.8 (9.4-11.9), NR=0	9.8 (9.0-11.9), NR=0
Smoking	43.9 (38.0-59.8), NR=4	25.0 (13.4-34.7), NR=2	30.0 (20.1-30.1), NR=2	16.7 (13.0-19.3), NR=1	50.2 (26.9-70.1), NR=10	19.4 (6.3-26.7), NR=11
Diabetes	14.5 (7.0-42.0), NR=4	14.6 (7.4-51.0), NR=2	10.0 (0.0-17.0), NR=2	12.0 (0.0-13.0), NR=1	12.2 (6.2-43.0), NR=11	7.8 (5.5-43.6), NR=11
SBP mean	135.2 (132.4-138.5), NR=11	135.0 (133.1-135.8), NR=6	139.9 (136.6-142.0), NR=3	140.3 (136.8-144.3), NR=2	136.8 (127.9-143.0), NR=10	130.3 (126.3-140.0), NR=10
SBP sd	18.6 (17.4-21.0), NR=11	19.6 (18.8-22.0), NR=6	21.1 (21.0-21.3), NR=3	21.9 (21.8-22.0), NR=2	19.4 (18.1-21.0), NR=10	20.8 (20.6-22.4), NR=10
HDL mean	49.5 (47.9-53.5), NR=4	58.0 (56.3-62.0), NR=2	49.8 (47.7-52.0), NR=3	58.4 (57.5-59.8), NR=2	50.4 (50.0-54.1), NR=11	59.7 (54.2-69.6), NR=11
HDL sd	13.9 (12.0-15.5), NR=4	15.7 (12.7-19.0), NR=2	13.1 (12.0-15.0), NR=3	16.3 (15.6-17.2), NR=2	14.4 (13.4-17.2), NR=11	16.2 (15.0-20.1), NR=11
Total cholesterol mean	226.9 (212.6-239.3), NR=4	234.0 (216.7-239.0), NR=2	225.4 (209.7-234.2), NR=3	242.2 (227.1-258.7), NR=2	217.0 (196.9-235.3), NR=10	224.3 (203.0-239.8), NR=10
Total cholesterol sd	40.2 (37.2-43.7), NR=4	42.0 (38.0-52.7), NR=2	37.7 (35.5-43.0), NR=3	39.3 (36.5-45.8), NR=2	37.4 (36.1-42.5), NR=10	40.1 (38.2-47.4), NR=10

Values indicate N (%), or median (IQR)
PCE: Pooled Cohort Equations, NR: not reported, CHD: coronary heart disease, CVD: cardiovascular disease, excl: excluded, sd: standard deviation, SBP: systolic blood pressure, HDL: high density lipoprotein cholesterol

3.4 Summary calibration slope

Table S3: Results of summary calibration slope

Model	Calibration slope	95% CI	95% PI
Wilson men	1.01	0.95-1.07	0.95-1.07
Wilson women	0.97	0.71-1.22	-0.06-2.00
ATP III men	1.29	0.97-1.82	0.14-2.45
ATP III women	0.95	Not estimable	0.87-1.03
PCE men	0.95	0.79-1.10	-0.19-2.07
PCE women	0.82	0.77-0.86	0.28-1.35

CI: confidence interval, PI: prediction interval

Meta-analysis of stratified OE ratios indicated that miscalibration of the Framingham models was mostly related to heterogeneity in baseline risk, as the summary calibration slope is close to 1. A calibration slope between 0 and 1 indicates predictions are too extreme, e.g. too low for low-risk people and too high for high-risk people. A calibration slope >1 indicates there is not enough variability in predicted risks.[55]

3.5 Sensitivity analyses

Table S4: Results of sensitivity analyses

		Wilson men		Wilson women		ATPIII men		ATPIII women		PCE men		PCE women
OE ratio	N	OE (95%CI)	N	OE (95%CI)	N	OE (95%CI)	N	OE (95%CI)	N	OE (95%CI)	N	OE (95%CI)
All validations	16	0.580 (0.434-0.726)	10	0.685 (0.442-0.928)	5	0.581 (0.368-0.793)	4	0.785 (0.596-0.974)	19	0.661 (0.591-0.731)	20	0.763 (0.646-0.881)
Low risk of bias for all domains*	1	-	1	-	4	-	1	-	2	-	3	-
Low risk of bias for participant selection, predictors and outcome	10	0.628 (0.440-0.817)	6	0.720 (0.343-1.097)	4	0.581 (0.335-0.827)	4	0.785 (0.596-0.974)	16	0.683 (0.607-0.760)	17	0.797 (0.676-0.918)
Weighted by number of events	16	0.580 (0.434-0.726)	10	0.685 (0.442-0.928)	5	0.557 (0.369-0.744)	4	0.784 (0.595-0.974)	19	0.660 (0.593-0.727)	20	0.781 (0.656-0.905)
Bivariate analyses	18	0.547 (0.384-0.384)	10	0.594 (0.387-0.91)	6	0.643 (0.44-0.94)	5	0.723 (0.559-0.936)	20	0.659 (0.596-0.728)	21	0.753 (0.645-0.878)
Not extrapolated to 10 year	16	0.575 (0.428-0.721)	10	0.676 (0.429-0.923)	5	0.581 (0.368-0.793)	4	0.785 (0.596-0.974)	19	0.657 (0.587-0.728)	20	0.763 (0.646-0.880)
C-statistic	N	C (95%CI)	N	C (95%CI)	N	C (95%CI)	N	C (95%CI)	N	C (95%CI)	N	C (95%CI)
All validations	18	0.676 (0.659-0.693)	10	0.706 (0.657-0.756)	5	0.636 (0.594-0.679)	4	0.660 (0.648-0.673)	20	0.701 (0.679-0.723)	20	0.741 (0.719-0.763)
Low risk of bias for all domains*	2	-	2	-	4	-	2	-	2	-	2	-

Table S4: Continued

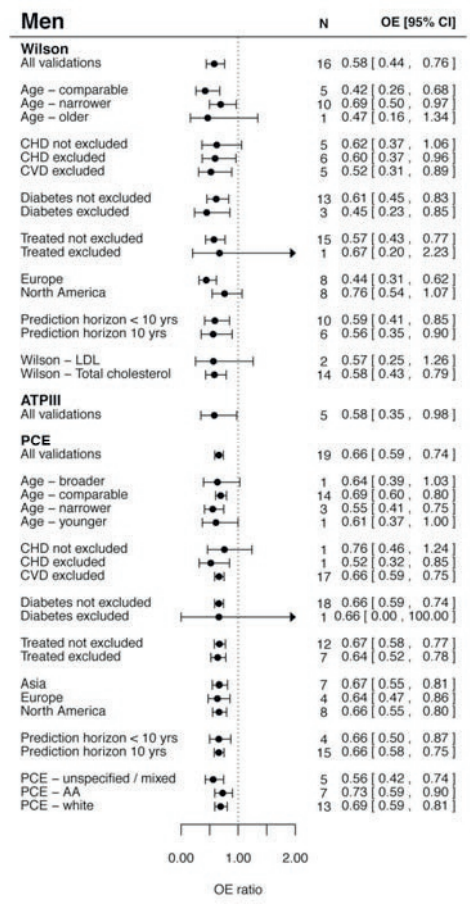
		Wilson men		Wilson women		ATPIII men		ATPIII women		PCE men		PCE women
Low risk of bias for participant selection, predictors and outcome	12	0.680 (0.659-0.702)	8	0.706 (0.647-0.766)	4	0.642 (0.588-0.696)	4	0.660 (0.648-0.673)	16	0.711 (0.689-0.734)	16	0.751 (0.729-0.774)
Weighted by number of events	18	0.675 (0.657-0.694)	10	0.690 (0.643-0.736)	5	0.638 (0.595-0.68)	4	0.658 (0.648-0.669)	20	0.701 (0.679-0.724)	20	0.742 (0.72-0.765)
Bivariate analyses	18	0.676 (0.660-0.691)	10	0.707 (0.655-0.754)	6	0.629 (0.588-0.668)	5	0.660 (0.640-0.680)	20	0.701 (0.677-0.724)	21	0.740 (0.718-0.761)

*No summary statistics are reported because of the low number of validations. OE: observed expected.

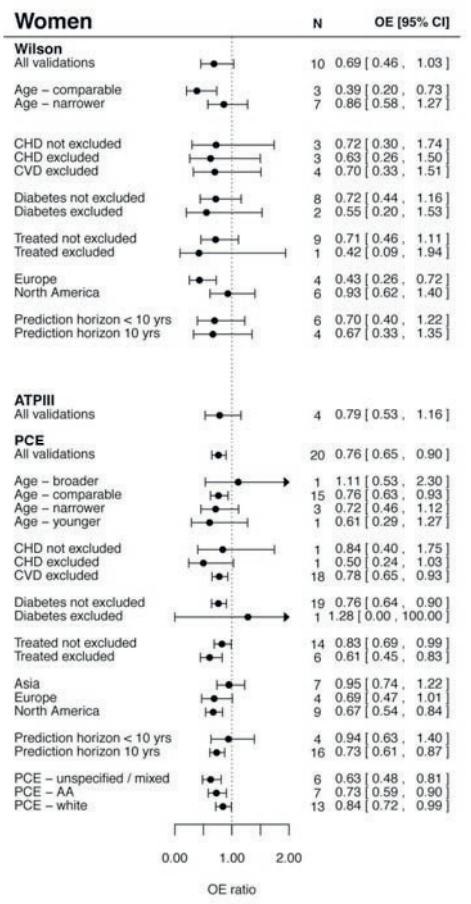
3.6 Metaregression analyses

3.6.1 OE ratio

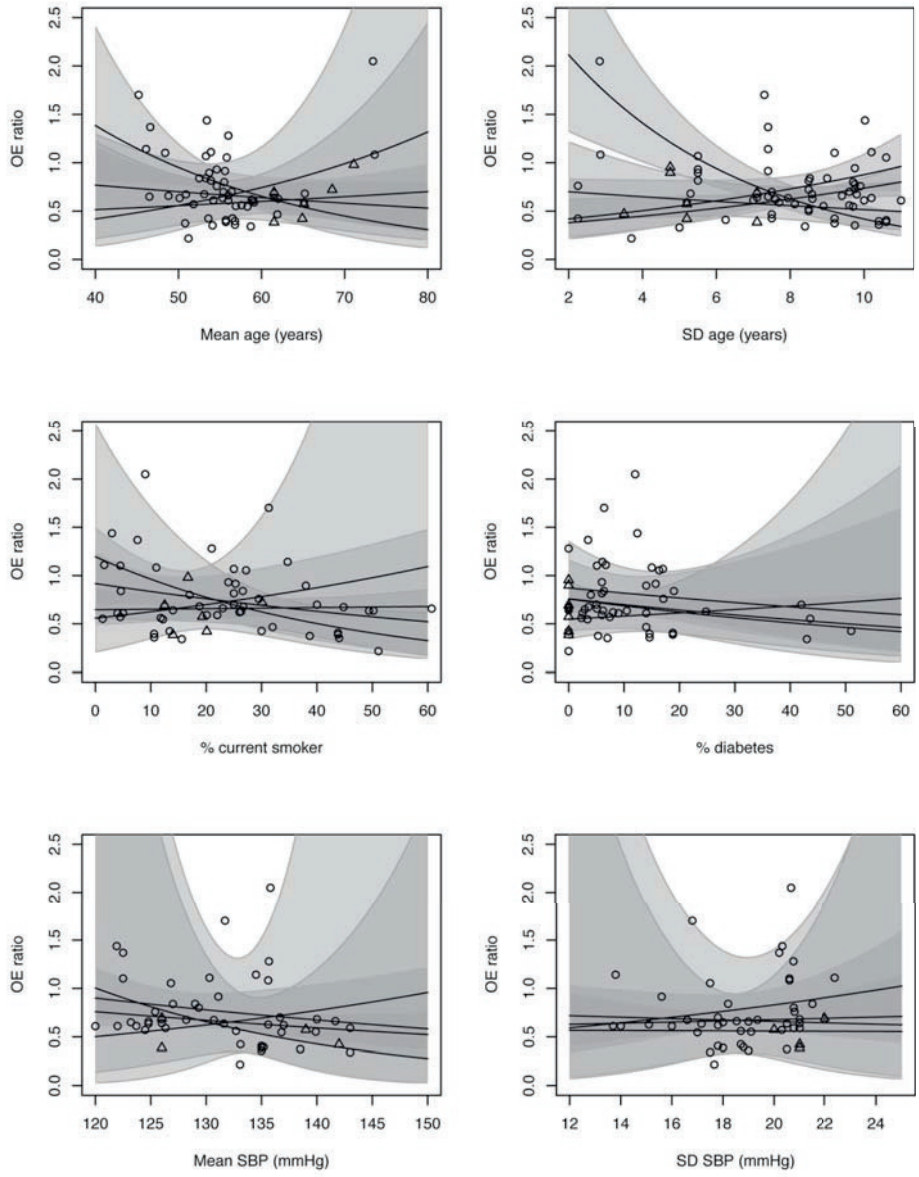
A



B



c


OE ratio
Mean age (years)
SD age (years)
% current smoker
% diabetes
Mean SBP (mmHg)
SD SBP (mmHg)

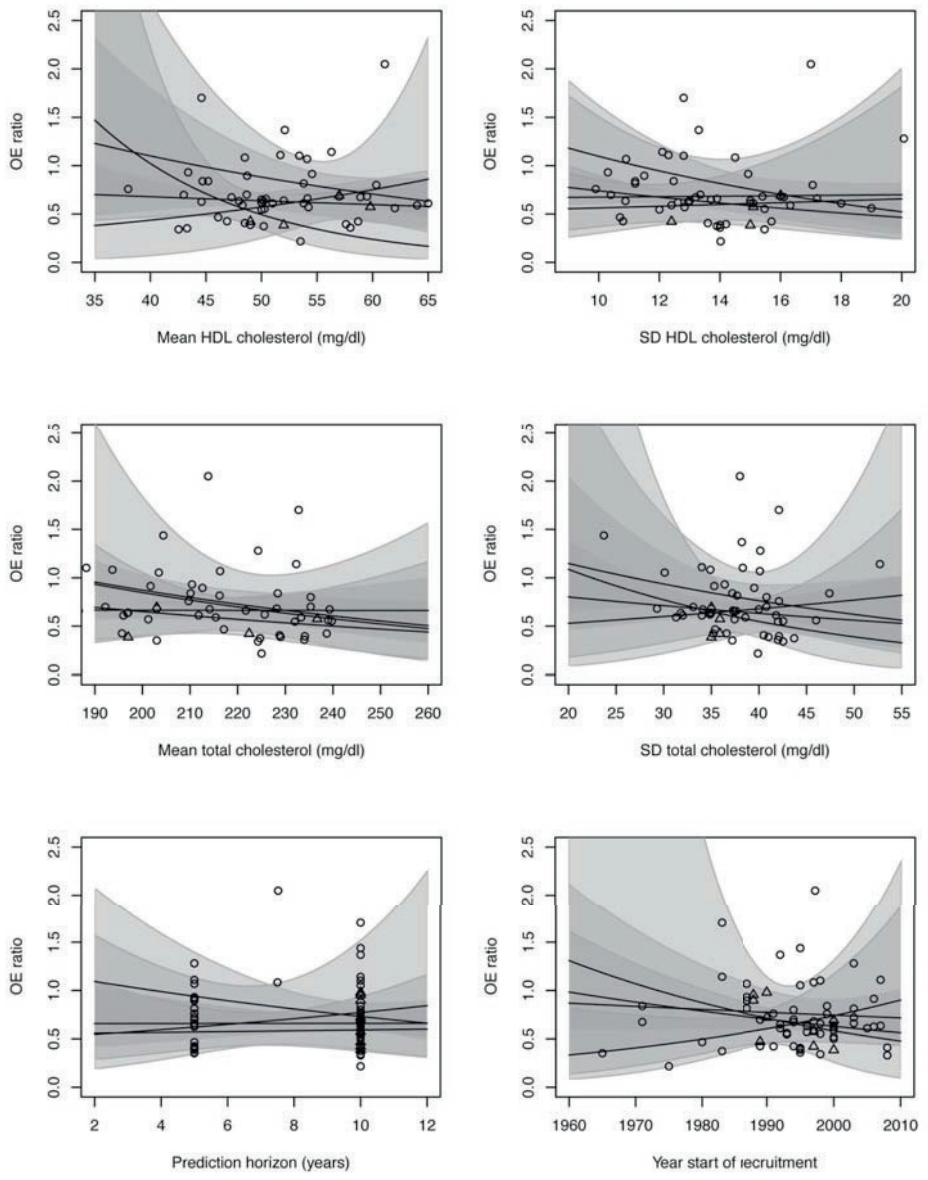
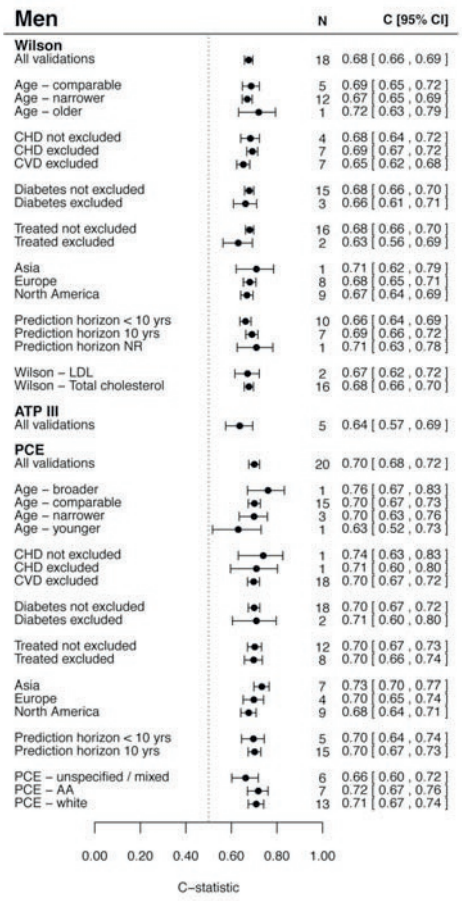


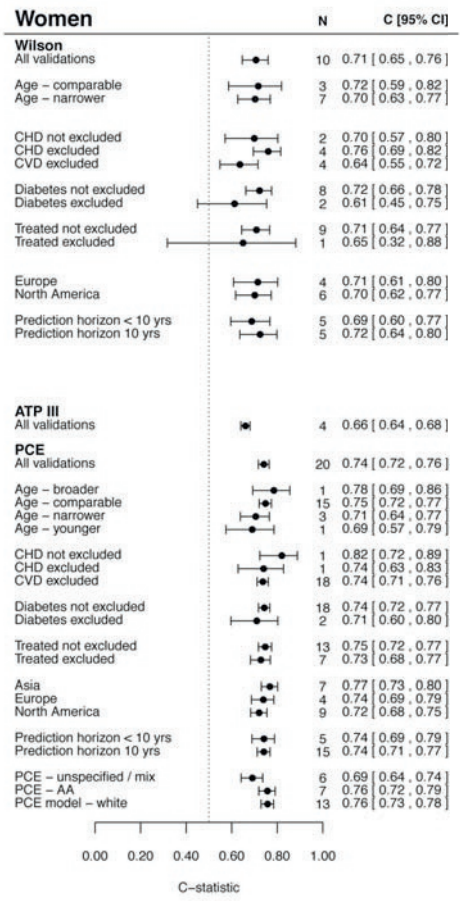
Figure S1: Results of meta-regression OE ratio for categorical variables (A and B) and continuous variables (C). For C, every line represents one model: Wilson men, Wilson women, PCE men or PCE women. ATP III is not plotted because of the low number of external validations, but the triangles represent the individual validations for the ATP III models. The grey areas represent the confidence intervals around the lines, and the circles represent the individual external validations. CHD: coronary heart disease, CVD: cardiovascular disease, AA: African American, SD: standard deviation, SBP: systolic blood pressure, HDL: high-density lipoprotein.

3.6.2 C-statistic

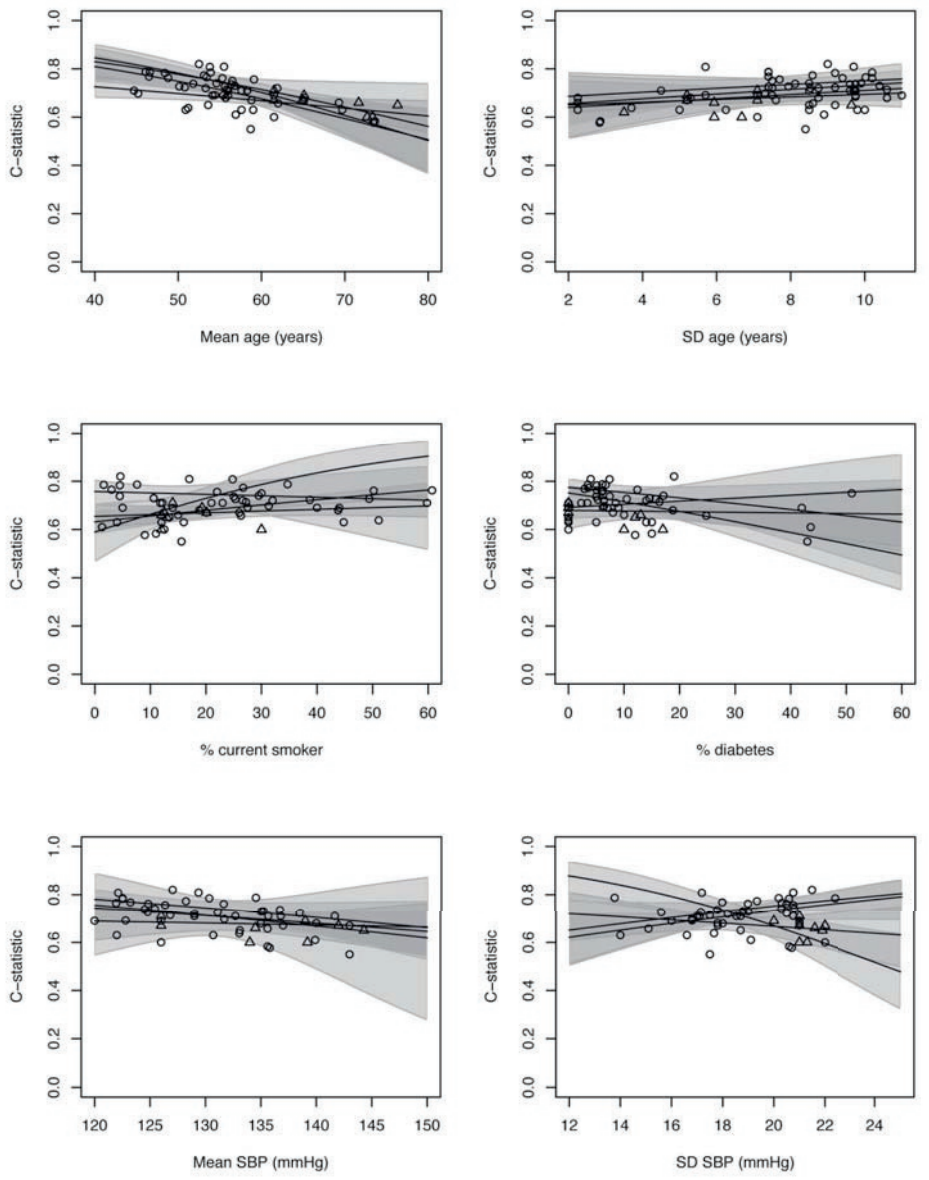
A



B



C


C–statistic
0.0
0.2
0.4
0.6
0.8
1.0
40
50
60
70
80
Mean age (years)
2
4
6
8
10
SD age (years)
0
10
20
30
40
50
60
% current smoker
% diabetes
120
125
130
135
140
145
150
Mean SBP (mmHg)
12
14
16
18
20
22
24
SD SBP (mmHg)

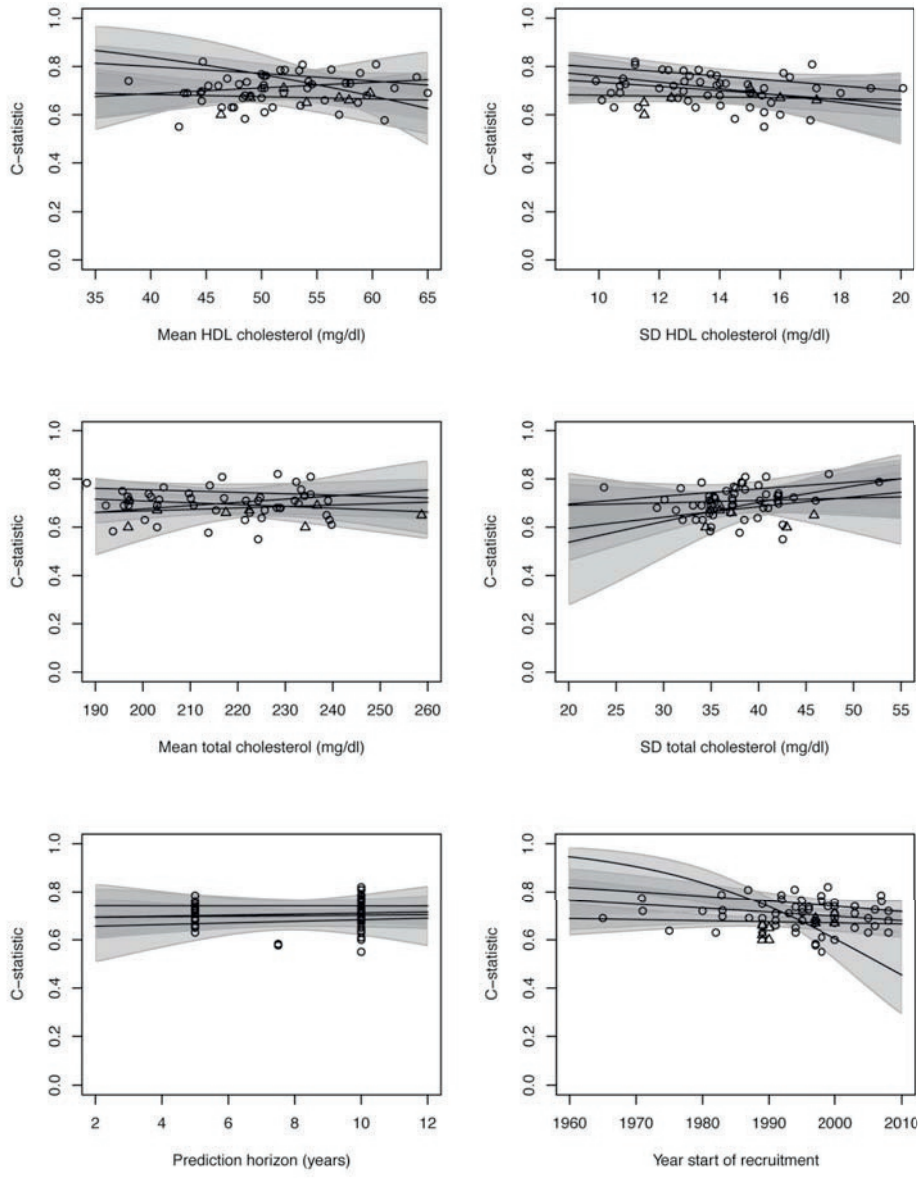


Figure S2: Results of meta-regression c-statistic for categorical variables (A and B) and continuous variables (C). For C, every line represents one model: Wilson men, Wilson women, PCE men or PCE women. ATP III is not plotted because of the low number of external validations, but the triangles represent the individual validations for the ATP III models. The grey areas represent the confidence intervals around the lines, and the circles represent the individual external validations. CHD: coronary heart disease, CVD: cardiovascular disease, AA: African American, SD: standard deviation, SBP: systolic blood pressure, HDL: high-density lipoprotein.

4 References

1. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97(18):1837-47.
2. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* 2002;106(25):3143-421.
3. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). *JAMA* 2001;285(19):2486-97.
4. Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014;129(25 Suppl 2):S49-73.
5. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol* 2014;14:135.
6. Mainous AG, 3rd, Koopman RJ, Diaz VA, Everett CJ, Wilson PWF, Tilley BC. A coronary heart disease risk score based on patient-reported information. *Am J Cardiol* 2007;99(9):1236-41.
7. Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: asymptotic methods and evaluation. *Stat Med* 2006;25(4):559-73.
8. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29-36.
9. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;356:i6460.
10. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* 2014;14:25.
11. Pennells L, Kaptoge S, White IR, Thompson SG, Wood AM. Assessing risk prediction models using individual participant data from multiple studies. *Am J Epidemiol* 2014;179(5):621-32.
12. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2015.

13. Lee J, Heng D, Ma S, Chew S-K, Hughes K, Tai ES. The metabolic syndrome and mortality: the Singapore Cardiovascular Cohort Study. *Clin Endocrinol (Oxf)* 2008;69(2):225-30.
14. Stork S, Feelders RA, van den Beld AW, Steyerberg EW, Savelkoul HFJ, Lamberts SWJ, et al. Prediction of mortality risk in the elderly. *Am J Med* 2006;119(6):519-25.
15. Barroso LC, Muro EC, Herrera ND, Ochoa GF, Hueros JIC, Buitrago F. Performance of the Framingham and SCORE cardiovascular risk prediction functions in a non-diabetic population of a Spanish health care centre: a validation study. *Scand J Prim Health Care* 2010;28(4):242-8.
16. Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. *JAMA* 2007;297(6):611-9.
17. Berry JD, Lloyd-Jones DM, Garside DB, Greenland P. Framingham risk score and prediction of coronary heart disease death in young men. *Am Heart J* 2007;154(1):80-6.
18. Dunder K, Lind L, Zethelius B, Berglund L, Lithell H. Evaluation of a scoring scheme, including proinsulin and the apolipoprotein B/apolipoprotein A1 ratio, for the risk of acute coronary events in middle-aged men: Uppsala Longitudinal Study of Adult Men (ULSAM). *Am Heart J* 2004;148(4):596-601.
19. Jung KJ, Jang Y, Oh DJ, Oh BH, Lee SH, Park SW, et al. The ACC/AHA 2013 pooled cohort equations compared to a Korean Risk Prediction Model for atherosclerotic cardiovascular disease. *Atherosclerosis* 2015;242(1):367-75.
20. DeFilippis AP, Young R, Carrubba CJ, McEvoy JW, Budoff MJ, Blumenthal RS, et al. An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort. *Ann Intern Med* 2015;162(4):266-75.
21. De Filippis AP, Young R, McEvoy JW, Michos ED, Sandfort V, Kronmal RA, et al. Risk score overestimation: The impact of individual cardiovascular risk factors and preventive therapies on the performance of the American Heart Association-American College of Cardiology-Atherosclerotic Cardiovascular Disease risk score in a modern multi-ethnic cohort. *Eur Heart J* 2017;38(8):598-608.
22. Muntner P, Colantonio LD, Cushman M, Goff DC, Jr., Howard G, Howard VJ, et al. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *JAMA* 2014;311(14):1406-15.
23. Yang X, Li J, Hu D, Chen J, Li Y, Huang J, et al. Predicting the 10-Year Risks of Atherosclerotic Cardiovascular Disease in Chinese Population: The China-PAR Project (Prediction for ASCVD Risk in China). *Circulation* 2016;134(19):1430-40.
24. Mortensen MB, Afzal S, Nordestgaard BG, Falk E. Primary Prevention With Statins: ACC/AHA Risk-Based Approach Versus Trial-Based Approaches to Guide Statin Therapy. *J Am Coll Cardiol* 2015;66(24):2699-709.

25. Mortensen MB, Nordestgaard BG, Afzal S, Falk E. ACC/AHA guidelines superior to ESC/EAS guidelines for primary prevention with statins in non-diabetic Europeans: the Copenhagen General Population Study. *Eur Heart J* 2017;38(8):586-94.
26. D'Agostino RB, Sr., Grundy S, Sullivan LM, Wilson P. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA* 2001;286(2):180-7.
27. Koller MT, Leening MJG, Wolbers M, Steyerberg EW, Hunink MGM, Schoop R, et al. Development and validation of a coronary risk prediction model for older U.S. and European persons in the cardiovascular health study and the Rotterdam Study. *Ann Intern Med* 2012;157(6):389-97.
28. Koller MT, Steyerberg EW, Wolbers M, Stijnen T, Bucher HC, Hunink MGM, et al. Validity of the Framingham point scores in the elderly: results from the Rotterdam study. *Am Heart J* 2007;154(1):87-93.
29. Andersson C, Enserro D, Larson MG, Xanthakis V, Vasan RS. Implications of the US cholesterol guidelines on eligibility for statin therapy in the community: comparison of observed and predicted risks in the Framingham Heart Study Offspring Cohort. *Journal of the American Heart Association* 2015;4(4).
30. Buitrago F, Calvo-Hueros JI, Canon-Barroso L, Pozuelos-Estrada G, Molina-Martinez L, Espigares-Arroyo M, et al. Original and REGICOR Framingham functions in a nondiabetic population of a Spanish health care center: a validation study. *Ann Fam Med* 2011;9(5):431-8.
31. Chia YC, Lim HM, Ching SM. Validation of the pooled cohort risk score in an Asian population - a retrospective cohort study. *BMC Cardiovasc Disord* 2014;14:163.
32. Comin E, Solanas P, Cabezas C, Subirana I, Ramos R, Gene-Badia J, et al. Estimating cardiovascular risk in Spain using different algorithms. *Rev Esp Cardiol* 2007;60(7):693-702.
33. Cook NR, Ridker PM. Further insight into the cardiovascular risk calculator: the roles of statins, revascularizations, and underascertainment in the Women's Health Study. *JAMA Intern Med* 2014;174(12):1964-71.
34. Cooper JA, Miller GJ, Humphries SE. A comparison of the PROCAM and Framingham point-scoring systems for estimation of individual risk of coronary heart disease in the Second Northwick Park Heart Study. *Atherosclerosis* 2005;181(1):93-100.
35. De Las Heras Gala T, Geisel MH, Peters A, Thorand B, Baumert J, Lehmann N, et al. Recalibration of the ACC/AHA risk score in two population-based German cohorts. *PLoS One* 2016;11 (10) (no pagination)(e0164688):e0164688.
36. Emdin CA, Khera AV, Natarajan P, Klarin D, Baber U, Mehran R, et al. Evaluation of the Pooled Cohort Equations for Prediction of Cardiovascular Risk in a Contemporary Prospective Cohort. *Am J Cardiol* 2017;119(6):881-85.

37. Empana JP, Ducimetiere P, Arveiler D, Ferrieres J, Evans A, Ruidavets JB, et al. Are the Framingham and PROCAM coronary heart disease risk functions applicable to different European populations? The PRIME Study. *Eur Heart J* 2003;24(21):1903-11.
38. Ferrario M, Chiodini P, Chambless LE, Cesana G, Vanuzzo D, Panico S, et al. Prediction of coronary events in a low incidence population. Assessing accuracy of the CUORE Cohort Study prediction equation. *Int J Epidemiol* 2005;34(2):413-21.
39. Jee SH, Jang Y, Oh DJ, Oh BH, Lee SH, Park SW, et al. A coronary heart disease prediction model: The Korean heart study. *BMJ Open* 2014;4(5).
40. Kavousi M, Leening MJ, Nanchen D, Greenland P, Graham IM, Steyerberg EW, et al. Comparison of application of the ACC/AHA guidelines, Adult Treatment Panel III guidelines, and European Society of Cardiology guidelines for cardiovascular disease prevention in a European cohort. *JAMA* 2014;311(14):1416-23.
41. Khalili D, Asgari S, Hadaegh F, Steyerberg EW, Rahimi K, Fahimfar N, et al. A new approach to test validity and clinical usefulness of the 2013 ACC/AHA guideline on statin therapy: A population-based study. *Int J Cardiol* 2015;184(1):587-94.
42. Lee CH, Woo YC, Lam JKY, Fong CHY, Cheung BMY, Lam KSL, et al. Validation of the Pooled Cohort equations in a long-term cohort study of Hong Kong Chinese. *J Clin Lipidol* 2015;9(5):640-46.
43. Lloyd-Jones DM, Wilson PWF, Larson MG, Beiser A, Leip EP, D'Agostino RB, et al. Framingham risk score and prediction of lifetime risk for coronary heart disease. *Am J Cardiol* 2004;94(1):20-4.
44. Marrugat J, Subirana I, Comin E, Cabezas C, Vila J, Elosua R, et al. Validity of an adaptation of the Framingham cardiovascular risk function: the VERIFICA Study. *J Epidemiol Community Health* 2007;61(1):40-7.
45. Pike MM, Decker PA, Larson NB, St Sauver JL, Takahashi PY, Roger VL, et al. Improvement in Cardiovascular Risk Prediction with Electronic Health Records. *J Cardiovasc Transl Res* 2016;9(3):214-22.
46. Rana JS, Tabada GH, Solomon MD, Lo JC, Jaffe MG, Sung SH, et al. Accuracy of the Atherosclerotic Cardiovascular Risk Equation in a Large Contemporary, Multiethnic Population. *J Am Coll Cardiol* 2016;67(18):2118-30.
47. Reissigova J, Zvarova J. The Framingham risk function underestimated absolute coronary heart disease risk in Czech men. *Methods Inf Med* 2007;46(1):43-9.
48. Rodondi N, Locatelli I, Aujesky D, Butler J, Vittinghoff E, Simonsick E, et al. Framingham risk score and alternatives for prediction of coronary heart disease in older adults. *PLoS One* 2012;7(3):e34287.
49. Ryckman EM, Summers RM, Liu J, Munoz del Rio A, Pickhardt PJ. Visceral fat quantification in asymptomatic adults using abdominal CT: is it predictive of future cardiac events? *Abdom Imaging* 2015;40(1):222-6.

50. Simmons RK, Sharp S, Boekholdt SM, Sargeant LA, Khaw K-T, Wareham NJ, 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? *Arch Intern Med* 2008;168(11):1209-16.
51. Simons LA, Simons J, Friedlander Y, McCallum J, Palaniappan L. Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study. *Med J Aust* 2003;178(3):113-6.
52. Suka M, Sugimori H, Yoshida K. Application of the updated Framingham risk score to Japanese men. *Hypertens Res* 2001;24(6):685-9.
53. Sussman JB, Wiitala WL, Zawistowski M, Hofer TP, Bentley D, Hayward RA. The Veterans Affairs Cardiac Risk Score: Recalibrating the Atherosclerotic Cardiovascular Disease Score for Applied Use. *Med Care* 2017;55(9):864-70.
54. Vaidya D, Yanek LR, Moy TF, Pearson TA, Becker LC, Becker DM. Incidence of coronary artery disease in siblings of patients with premature coronary artery disease: 10 years of follow-up. *Am J Cardiol* 2007;100(9):1410-5.
55. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2014.

Chapter 5

Prediction of 10-year risk of coronary heart disease in the general population: incremental value of blood biomarkers over traditional predictors in a pan-European cohort study

Johanna AAG Damen*
Linda M Peelen*
Romin Pajouheshnia
Camille M Lassale
Yvonne T van der Schouw
Ewoud Schuit
Ioanna Tzoulaki
Karel GM Moons
on behalf of the EPIC-CVD Consortium

*Authors equally contributed

Manuscript in preparation

Abstract

Background: Predictive ability of prediction models for future risk of cardiovascular disease (CVD) is suboptimal. The aim of this study is to assess the incremental value of candidate biomarkers above traditional predictors for the prediction of 10-year risk of coronary heart disease (CHD) in the general population.

Methods: The EPIC-CVD case-cohort study consists of 12261 men (6653 CHD events) and 14366 women (4484 CHD events) from 10 European countries. Prentice weighted Cox proportional hazards models were fitted adding the following plasma biomarkers to a traditional prediction models, separately: non-HDL cholesterol, triglycerides, apolipoprotein (apo) A1, apoB, lipoprotein(a) (Lp(a)), C-reactive protein (CRP), albumin, creatinine, uric acid, glucose, glycated hemoglobin (HbA1c), alkaline phosphatase (ALP), alanine transaminase (ALT), aspartate aminotransferase (AST), gamma-glutamyl transferase, bilirubin, calcium, magnesium, iron, and ferritin. Improvement in predictive performance was assessed in terms of discrimination (c-statistic), calibration (total observed expected (OE) ratio and calibration plots), and reclassification (net reclassification improvement (NRI)), at predefined probability thresholds.

Results: Median follow-up time was 10.8y (IQR 7.4-13.3). In males, a model including non-HDL cholesterol, triglycerides, apoA1, apoB, CRP, albumin, creatinine, uric acid, glucose, HbA1c, ALP, and iron improved predictive performance compared to the traditional prediction model (c-statistic 0.740 (95% CI 0.730-0.750) vs. 0. 728 (95% CI 0.718-0.738); OE ratio 0.995 (95% CI 0.968-1.021) vs. 1.010 (95% CI 0.983-1.036); NRI cases 0.048, non-cases -0.003). In females, a model including apoB, Lp(a), CRP, albumin, glucose, HbA1c, ALT, AST, and magnesium, showed added predictive value (c-statistic 0.771 (95% CI 0.762-0.781) vs. 0. 763 (95% CI 0.753-0.772); OE ratio 1.116 (95% CI 1.080-1.153) vs. 1.116 (95% CI 1.080-1.152); NRI cases 0.026, non-cases -0.001).

Conclusion: We identified several biomarkers that improved the prediction of 10-year CHD risk, albeit marginally. Additional impact analyses, including cost-effectiveness, are needed to determine whether this improvement in predictive performance as compared to using only the traditional predictors, indeed improves therapeutic decision making and subsequent patient outcomes.

Introduction

Cardiovascular disease (CVD), including coronary heart disease (CHD) and stroke, is a major public health problem accounting for 3.9 million deaths every year in Europe.[1] Established predictors of CVD include age, smoking, high body mass index (BMI), diabetes, hypertension, and hyperlipidemia.[2-5] Many of these are modifiable, providing the opportunity to prevent or delay CVD by means of lifestyle interventions, antihypertensive drugs or lipid lowering drugs, for example.

Plethora of prediction models have been developed to estimate the risk of having a CVD event in the coming years (e.g. within 10 years).[6] These models are being used to identify people at increased risk of future CVD events to guide prevention and target risk lowering interventions. Well-known examples of these models, which are recommended to use in clinical practice, are the Framingham risk scores,[7-9] the European Systematic COronary Risk Evaluation (SCORE),[4] QRISK,[5] and the ACC/AHA Pooled Cohort Equations.[2] Unfortunately, when applied to different populations, these prediction models often show poor predictive performance, i.e. models typically overestimate the risk of CVD[10,11] and many CVD events occur in people not classified as high risk by prediction models.[12] Since interventions, including treatment decisions, are advocated to be based on risk estimates in many clinical guidelines, incorrect predicted risks can result in over- or undertreatment. One strategy to improve the predictive performance of these models is to add more predictors. Additional predictors may result in more accurate predictions in certain individuals or better distinguish between people who will or will not develop the event of interest.[13,14]

Several biomarkers have been suggested to improve prediction of future CVD events, on top of the established predictors mentioned above. These include lipids (e.g. apolipoproteins),[15-18] markers of insulin resistance (e.g. HbA1c),[19] liver enzymes (e.g. alkaline phosphatase),[20] iron parameters,[21,22] uric acid,[23] and inflammatory markers (e.g. C-reactive protein (CRP)).[24,25] The existing literature is, however, fragmented with no real head-to-head comparisons of biomarkers and no studies examining all markers together.

This study aimed to investigate the incremental value of such biomarkers beyond the traditional CVD predictors in predicting 10-year risk of CHD in a multicentre pan-European cohort of over 26000 individuals.

Methods

This paper is written according to the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines (Table S1).[26,27]

Study population

The European Prospective Investigation into Cancer and Nutrition (EPIC) study is a large multicentre cohort study consisting of 519,978 adults from 29 centres across 10 European countries (Norway, Sweden, Denmark, the Netherlands, the United Kingdom, Germany, France, Italy, Spain and Greece).[28] Participants were recruited from the general population between 1991 and 1999. Eligibility criteria and follow-up time per centre are listed in Table S2. At baseline, information was collected for all participants on diet, lifestyle characteristics, anthropometric measurements, and medical history, and blood samples were taken and stored. EPIC-CVD is a case-cohort study within the EPIC study, focusing on development of CVD.[29] EPIC-CVD consists of a randomly selected subcohort of 16,242 participants who had available stored blood and buffy coat, supplied with all cases of CHD that occurred outside the subcohort during follow-up. Participants with a history of myocardial infarction or stroke were excluded. EPIC-CVD has a median follow-up of 10.8 years (IQR 7.4-13.3).

Traditional predictors

The traditional predictors used in this study consisted of predictors included in the majority of prediction models for CVD,[6] namely age, current smoking, diabetes (self-reported), BMI, systolic blood pressure (SBP), hypertension (self-reported hypertension, self-reported use of anti-hypertensive medication, SBP > 140mmHg, and/or diastolic blood pressure (DBP) > 90mmHg), total cholesterol, and high density lipoprotein (HDL) cholesterol. Traditional predictors were all recorded at baseline.

Biomarkers

Biomarkers considered for extending the traditional prediction model were non-HDL cholesterol, triglycerides, apolipoprotein A1 (apoA1), apolipoprotein B (apoB), lipoprotein(a) (Lp(a)), C-reactive protein (CRP), albumin, creatinine, uric acid, glucose, glycated hemoglobin (HbA1c), alkaline phosphatase (ALP), alanine transaminase (ALT), aspartate aminotransferase (AST), gamma-glutamyl transferase (GGT), total bilirubin, calcium, magnesium, iron, and ferritin. All biomarkers were measured in baseline serum samples at Stichting Huisartsen Laboratorium (Etten-Leur, the Netherlands) using a Cobas enzymatic assay (Roche Diagnostics, Mannheim, Germany) on a Roche Hitachi Modular P analyser, except for HbA1c, which was measured in erythrocytes using the Tosoh-G8 HPLC analyser (Tosoh Bioscience, Japan). Transferrin was not measured directly but calculated as half of the total iron binding capacity, and non-HDL cholesterol was calculated as total cholesterol minus HDL-cholesterol.

Outcome

Follow-up for outcomes was done as part of the EPIC study. Our main outcome is the occurrence of CHD within 10 years. CHD was defined as fatal or nonfatal myocardial

infarction, angina, and other types of acute or chronic coronary heart diseases, with ICD-10 codes I20-I25. Fatal events that occurred within 28 days of a nonfatal event, were considered as a single fatal event. Nonfatal events were recorded using follow-up questionnaires or linkage with morbidity or hospital registries, and death registries were used for fatal events. Suspected events were validated using medical records, contact with clinicians, death certificates or contact with relatives of deceased participants. Participants were censored if they had a CHD event, died from a non-CHD cause, were lost to follow-up, or reached the end of the follow-up period (Table S2). All outcomes were assessed blinded for the predictors.

Ethics committee and informed consent

The EPIC study complies with the Declaration of Helsinki, and all participants gave written informed consent. The study was approved by the local ethics committees of the participating centres and the Internal Review Board of the International Agency for Research on Cancer (IARC, Lyon).

Statistical analyses

The goal of our analyses was to determine the incremental predictive value of biomarkers beyond the traditional predictors, and not to develop a new prediction model. All analyses were conducted separately for males and females.

Baseline characteristics were described for cases and non-cases separately, and by country. We fitted Prentice weighted[30] Cox proportional hazards models to account for the case-cohort design, using the full available follow-up (i.e. events happening after 10 years were considered a case), and used these models to predict the risk of CHD at 10 years. Models were stratified by country to account for clustering and differences in baseline risk, i.e. baseline hazards were allowed to vary between countries, but predictor weights were the same for all countries. Within the United Kingdom cohorts, the Oxford centre was analysed separately because of the characteristics of its population (mostly vegetarian, health-conscious participants).[28] Transformations of continuous predictors were determined by assessing Martingale residuals. We verified the need for transformation of continuous variables by including fractional polynomials or restricted cubic splines,[31] and subsequently comparing model fit (in terms of Bayesian Information Criterion (BIC)). We checked the proportional hazards assumption for the full model using Schoenfeld residuals adjusted for the case-cohort design[32] and log-log plots (for categorical variables).

We performed our analyses in a stepwise approach. First, we performed a univariable analysis to determine the association between each of the 29 predictors (traditional predictors and biomarkers) and the outcome. Second, we fitted the model with the traditional predictors only (further referred to as 'traditional prediction model'). In the third step, we fitted models in which we added each of the biomarkers separately to the

traditional prediction model (21 models in total). In the fourth step, we fitted a model in which we added all biomarkers to the traditional prediction model to assess the maximal predictive performance in this dataset. Finally, to investigate the potential performance of the model including all biomarkers with most incremental value in previous steps, we fitted a combined model, with traditional predictors, and the biomarkers that had incremental value in the third step, with incremental value defined as described in the following section. Models from step 3-5 will be further referred to as 'extended models'.

Model performance

The apparent predictive performance of the extended models was expressed in terms of discrimination, calibration, reclassification as compared to the traditional prediction model, R^2, Akaike Information Criterion (AIC), BIC, and Brier score. To assess the discriminative ability of the model, we calculated a weighted version of Harrell's concordance (c)-statistic, to account for the case-cohort design.[33] The c-statistic ranges between 0 and 1, where 1 means perfect discriminative ability, while 0.5 indicates the model is not better than chance.[34] For calibration we accounted for the case-cohort design by weighting every observation in the dataset by the sampling fraction for the respective centre. We calculated the ratio between the number of observed (O) cases over the number of expected (E) cases as predicted by the model (OE ratio). Furthermore, 10-year risk calibration plots were made in which observed risks were plotted against predicted risks, in deciles of predicted risk. Reclassification tables were created for every extended model, as compared to the traditional prediction model, in 4 categories of predicted risk: 0-5%, 5-10%, 10-20% and >20%.[4,35] Based on these tables we calculated the categorical net reclassification index (NRI), separate for cases and non-cases. To account for the case-cohort design, cases outside the subcohort that experienced a CHD event after 10 years follow-up were excluded from the NRI calculation.[33] As a sensitivity analysis, the NRI was also calculated for the categories 0-5%, 5-7.5%, 7.5-15%, >15%, and 0-7.5%, 7.5-15%, >15%.[2]

Selection of biomarkers for the combined model was based on the differences in c-statistic and OE ratio between the extended models from step 3 and the traditional prediction model, and the categorical NRIs. We selected the five biomarkers with highest incremental value for any of these performance measures for inclusion in the combined model. R^2, AIC, BIC and Brier score were used to verify that no important predictors were overlooked by this selection strategy (i.e. biomarkers that were not selected based on c-statistic, OE ratio, or NRI, but did have significant added value on these other performance measures). If the Spearman correlation between two biomarkers was >0.7, only the clinically most relevant (e.g. routinely tested in clinical practice) biomarker was retained, or the one with the highest increase in performance after adding it to the traditional prediction model if no such comparison of relevance was possible.

Model performance stratified by country

To study differences between countries, we calculated the performance of models separately for every country, i.e. models were fitted as described before with stratified baseline hazards per country, and these models were then applied to every country separately. For the traditional prediction model with every biomarker added separately (step 3), we calculated the c-statistics and OE ratios, to assess incremental value of the biomarkers in the different countries.

Missing data

Baseline characteristics of participants with and without missing predictor values were compared. Missing values were imputed multiple times with chained equations, using the 'mice' package in R.[36] We included all traditional predictors, biomarkers, and the outcome, and added variables regarding country, socioeconomic status, physical activity, alcohol use, event status, and the Nelson-Aalen estimator of the baseline cumulative hazard to impute missing values.[37] Ten imputations were performed with a maximum of 150 iterations. Correlated variables (>0.5) were identified and one was selected based on the correlation between that variable and the variable that needed imputation. All analyses described above were performed in each imputation set separately and combined using Rubin's rule.[38] C-statistics, OE ratios and hazard ratios (HR) were pooled with Rubin's rules, after logit (c-statistic) or log (OE ratio and HR) transformation.[39] For NRI, Brier score, AIC, BIC, and R^2 we reported the median of the 10 imputations. Selection of predictors for the combined model was done after pooling, i.e. the process of predictor selection was not repeated in every imputed dataset. All analyses were performed in R version 3.3.2.[40]

Results

Baseline characteristics

The subcohort consisted of 5946 men and 10087 women, of which 338 (5.7%) and 205 (2.0%) were CHD cases, respectively. 6315 male and 4279 female CHD cases were added to the dataset from outside the subcohort. The characteristics of the included participants with complete data for all predictors are described in Table 1 for the full cohort and by country in Table S3. Baseline characteristics of participants with and without missing data were comparable (Table S4). Median follow-up time was 9.9 years (IQR 6.2-13.1) for males and 11.3 years (IQR 8.3-13.4) for females.

Associations with cardiovascular risk

Table 2 shows the associations between each predictor and CHD within 10 years. From the univariable analyses it can be concluded that all predictors except for uric acid, ALP

(males), ALT (males), AST, calcium (males), magnesium, iron and transferrin (males) were associated with the occurrence of CHD. When added to the traditional prediction model, increased levels of non-HDL cholesterol (males), apoB, Lp(a), CRP, creatinine (males), glucose (males), HbA1c, and decreased levels of triglycerides (males), albumin, calcium (males) were statistically significantly associated with a higher risk of CHD. Proportional hazards assumptions were met for most, but not all, variables in the model. The use of fractional polynomials or restricted cubic splines was challenging with respect to model convergence and did not improve model fit.

Table 1: baseline characteristics of study participants. Separate for males and females and cases and non-cases.

	Males				Females			
	No CHD		CHD		No CHD		CHD	
		Missing		Missing		Missing		Missing
N	3307		3931		6524		2937	
Age (years)	53.6 (8.6)	0 (0.0%)	58.5 (7.9)	0 (0.0%)	53.5 (8.9)	0 (0.0%)	60.0 (7.6)	0 (0.0%)
Current smoker	1023 (30.9%)	42 (0.7%)	1530 (38.9%)	25 (0.6%)	1435 (22.0%)	63 (0.4%)	900 (30.6%)	15 (0.9%)
BMI (kg/m2)	26.4 (3.5)	0 (0.0%)	27.0 (3.6)	0 (0.0%)	25.6 (4.5)	0 (0.0%)	26.7 (4.6)	0 (0.0%)
Diabetes	103 (3.1%)	0 (0.0%)	281 (7.1%)	0 (0.0%)	124 (1.9%)	0 (0.0%)	218 (7.4%)	0 (0.0%)
Hypertension	1458 (44.1%)	0 (0.0%)	2501 (63.6%)	0 (0.0%)	2450 (37.6%)	0 (0.0%)	1894 (64.5%)	0 (0.0%)
SBP (mmHg)	135.3 (18.3)	1465 (25.9%)	143.7 (20.3)	1209 (22.0%)	129.9 (19.8)	2170 (14.8%)	142.6 (22.3)	630 (32.7%)
Use of antihypertensive medication at baseline	434 (13.1%)	3958 (70.0%)	932 (23.7%)	6470 (40.0%)	670 (10.3%)	3748 (59.5%)	770 (26.2%)	1972 (88.3%)
Use of lipid lowering medication at baseline	105 (3.2%)	4207 (74.4%)	200 (5.1%)	7859 (42.5%)	159 (2.4%)	4150 (63.2%)	178 (6.1%)	3081 (93.8%)
Total cholesterol (mmol/l)	5.9 (1.1)	299 (5.3%)	6.3 (1.1)	1248 (4.5%)	6.0 (1.1)	357 (3.0%)	6.6 (1.2)	947 (6.7%)
HDL cholesterol (mmol/l)	1.3 (0.4)	299 (5.3%)	1.2 (0.3)	1249 (4.5%)	1.6 (0.4)	357 (3.0%)	1.5 (0.4)	951 (6.7%)
non-HDL cholesterol (mmol/l)	4.6 (1.1)	300 (5.3%)	5.1 (1.1)	1249 (4.5%)	4.4 (1.2)	360 (3.0%)	5.2 (1.3)	951 (6.7%)
Triglycerides (mmol/l)	1.3 (1.0-1.9)	298 (5.3%)	1.7 (1.2-2.3)	1249 (4.5%)	1.0 (0.8-1.5)	361 (3.0%)	1.4 (1.0-2.0)	947 (6.6%)
ApoA1 (g/l)	1.4 (0.2)	299 (5.3%)	1.4 (0.2)	1249 (4.5%)	1.6 (0.3)	358 (3.0%)	1.6 (0.3)	947 (6.7%)
ApoB (g/l)	1.0 (0.2)	299 (5.3%)	1.2 (0.3)	1252 (4.5%)	1.0 (0.3)	366 (3.0%)	1.2 (0.3)	950 (6.7%)

Table 1: Continued

	Males				Females			
	No CHD		CHD		No CHD		CHD	
		Missing		Missing		Missing		Missing
LP(a) (mg/dl)	41.4 (22.5-70.3)	462 (8.2%)	50.2 (28.4-83.3)	1481 (6.9%)	34.9 (18.3-62.4)	643 (4.7%)	49.4 (25.9-94.0)	1074 (10.3%)
CRP (mg/l)	1.0 (0.6-2.1)	303 (5.4%)	1.7 (0.8-3.6)	1252 (4.6%)	1.1 (0.5-2.5)	360 (3.1%)	1.9 (0.9-4.1)	947 (6.8%)
Albumin (g/l)	46.4 (3.2)	400 (7.1%)	45.9 (3.2)	1408 (6.0%)	45.8 (3.1)	540 (4.0%)	45.6 (3.3)	1021 (8.9%)
Creatinine (umol/l)	78.0 (71.0-86.0)	373 (6.6%)	80.0 (72.0-90.0)	1392 (5.6%)	63.0 (56.0-70.0)	510 (3.8%)	65.0 (58.0-72.0)	1015 (8.3%)
Uric acid (umol/l)	336.4 (71.2)	373 (6.6%)	349.2 (75.4)	1386 (5.6%)	252.0 (62.8)	509 (3.8%)	279.9 (72.9)	1014 (8.3%)
Glucose (mmol/l)	5.1 (4.5-5.7)	773 (13.7%)	5.1 (4.3-5.9)	1568 (11.6%)	4.8 (4.1-5.3)	993 (7.8%)	4.9 (4.0-5.6)	1061 (17.2%)
HbA1c (%)	5.4 (5.3-5.7)	153 (2.7%)	5.6 (5.3-5.9)	1181 (2.3%)	5.4 (5.3-5.7)	172 (1.5%)	5.6 (5.3-6.0)	921 (3.4%)
ALP(iU/l)	64.0 (55.0-76.0)	777 (13.7%)	70.0 (59.0-83.0)	1556 (11.7%)	62.0 (50.0-76.0)	983 (7.9%)	73.0 (60.0-87.0)	1052 (17.3%)
ALT (iU/l)	22.0 (17.0-31.0)	304 (5.4%)	23.0 (17.0-30.0)	1259 (4.6%)	17.0 (13.0-22.0)	383 (3.1%)	18.0 (14.0-24.0)	957 (6.8%)
AST (iU/l)	29.0 (25.0-34.0)	376 (6.6%)	30.0 (26.0-35.0)	1383 (5.7%)	26.0 (22.0-30.0)	513 (3.8%)	27.0 (24.0-32.0)	1017 (8.4%)
GGT (iU/l)	28.0 (20.0-44.0)	299 (5.3%)	31.0 (22.0-47.0)	1248 (4.5%)	17.0 (13.0-24.0)	367 (3.0%)	20.0 (15.0-31.0)	951 (6.7%)
Calcium (mmol/l)	2.4 (0.1)	787 (13.9%)	2.4 (0.1)	1559 (11.8%)	2.4 (0.1)	993 (8.0%)	2.5 (0.1)	1055 (17.6%)
Magnesium (mmol/l)	0.8 (0.1)	768 (13.6%)	0.9 (0.1)	1545 (11.5%)	0.8 (0.1)	968 (7.8%)	0.8 (0.1)	1056 (17.1%)

Table 1: Continued

	Males				Females			
	No CHD		CHD		No CHD		CHD	
		Missing		Missing		Missing		Missing
Iron (umol/l)	18.1 (5.8)	772 (13.6%)	17.8 (5.8)	1549 (11.6%)	16.7 (6.0)	974 (7.8%)	16.8 (5.6)	1050 (17.2%)
Transferrin (umol/l)	33.3 (4.8)	788 (13.9%)	33.4 (4.8)	1563 (11.8%)	34.9 (5.5)	997 (8.0%)	34.5 (5.2)	1056 (17.6%)
Ferritin (pmol/l)	337.1 (188.8-556.2)	388 (6.9%)	332.6 (186.5-548.3)	1397 (5.8%)	148.3 (74.2-260.7)	532 (3.9%)	206.7 (119.1-328.1)	1017 (8.7%)
Total bilirubin (umol/l)	8.0 (6.0-11.0)	429 (7.6%)	7.0 (5.0-10.0)	1485 (6.4%)	6.0 (5.0-9.0)	565 (4.3%)	6.0 (4.0-8.0)	1061 (9.6%)
France	0 (0.0%)		0 (0.0%)		510 (7.8%)		36 (1.2%)	
Italy	573 (17.3%)		432 (11.0%)		1150 (17.6%)		354 (12.1%)	
Spain	154 (4.7%)		90 (2.3%)		354 (5.4%)		55 (1.9%)	
UK	306 (9.3%)		1006 (25.6%)		455 (7.0%)		562 (19.1%)	
Netherlands	147 (4.4%)		347 (8.8%)		959 (14.7%)		840 (28.6%)	
Greece	374 (11.3%)		188 (4.8%)		622 (9.5%)		84 (2.9%)	
Germany	496 (15.0%)		214 (5.4%)		785 (12.0%)		83 (2.8%)	
Sweden	436 (13.2%)		662 (16.8%)		814 (12.5%)		423 (14.4%)	
Denmark	790 (23.9%)		915 (23.3%)		756 (11.6%)		366 (12.5%)	
Oxford*	31 (0.9%)		77 (2.0%)		119 (1.8%)		134 (4.6%)	

Values represent N (%), mean (standard deviation), or median (25^{th} - 75^{th} percentile). CHD: coronary heart disease, SBP: systolic blood pressure, HDL: high-density lipoprotein, apo: apolipoprotein, Lp(a): lipoprotein(a), CRP: C-reactive protein, HbA1c: glycated hemoglobin, ALP: alkaline phosphatase, ALT: alanine transaminase, AST: aspartate aminotransferase, GGT: gamma-glutamyl transferase. *Oxford treated as separate country because of large difference with other UK cohorts.

Table 2: HRs of univariable analyses, traditional prediction model, combined model and full model for all predictors, separate for males and females.

	Males					Females				
	Univariable	**TP**	**TP + biomarker**	**Full**	**Combined**	**Univariable**	**TP**	**TP + biomarker**	**Full**	**Combined**
Age	1.06 (1.06-1.07)	1.06 (1.05-1.06)		1.05 (1.04-1.05)	1.05 (1.04-1.05)	1.09 (1.09-1.10)	1.07 (1.06-1.08)		1.06 (1.05-1.07)	1.07 (1.06-1.07)
Current smoking	1.72 (1.58-1.88)	1.75 (1.58-1.94)		1.60 (1.42-1.79)	1.58 (1.41-1.77)	1.54 (1.41-1.69)	2.15 (1.91-2.42)		2.06 (1.82-2.34)	2.06 (1.82-2.34)
Diabetes	2.38 (1.96-2.88)	1.55 (1.24-1.94)		1.07 (0.81-1.41)	1.06 (0.81-1.40)	5.00 (4.05-6.18)	2.45 (1.91-3.14)		1.62 (1.18-2.23)	1.66 (1.20-2.28)
Hypertension	2.13 (1.96-2.32)	1.26 (1.11-1.42)		1.25 (1.09-1.42)	1.25 (1.10-1.43)	3.07 (2.81-3.36)	1.47 (1.28-1.70)		1.49 (1.29-1.72)	1.51 (1.31-1.74)
Log bmi	7.87 (5.61-11.05)	1.59 (1.05-2.41)		1.15 (0.73-1.80)	1.17 (0.75-1.83)	8.43 (6.53-10.88)	2.06 (1.45-2.92)		1.13 (0.76-1.68)	1.19 (0.81-1.76)
SBP	1.02 (1.02-1.02)	1.01 (1.00-1.01)		1.01 (1.00-1.01)	1.01 (1.00-1.01)	1.03 (1.02-1.03)	1.01 (1.01-1.01)		1.01 (1.01-1.01)	1.01 (1.01-1.01)
Log total cholesterol	7.89 (5.95-10.47)	6.57 (4.85-8.89)		0.35 (0.04-3.29)	0.98 (0.39-2.44)	11.99 (9.20-15.63)	3.78 (2.76-5.17)		2.39 (0.30-19.03)	1.09 (0.46-2.58)
Log HDL cholesterol	0.30 (0.25-0.35)	0.31 (0.25-0.39)		0.51 (0.25-1.06)	0.45 (0.26-0.79)	0.23 (0.20-0.27)	0.38 (0.31-0.46)		0.56 (0.26-1.21)	0.63 (0.47-0.86)
Non-HDL cholesterol	1.46 (1.40-1.53)		1.57 (1.16-2.12)	1.20 (0.83-1.73)		1.45 (1.38-1.53)		1.02 (0.77-1.36)	0.86 (0.63-1.18)	
Log triglycerides	1.88 (1.74-2.04)		0.82 (0.72-0.93)	0.88 (0.74-1.04)	0.91 (0.78-1.07)	2.71 (2.46-2.99)		0.94 (0.80-1.12)	0.96 (0.75-1.22)	

Table 2: Continued

	Males				Females			
	Univariable	**TP**	**TP + biomarker**	**Full**	**Combined**	**Univariable**	**TP**	**TP + biomarker**	**Full**	**Combined**
ApoA1	0.50 (0.41-0.60)		0.69 (0.46-1.04)	1.17 (0.71-1.95)	1.08 (0.66-1.77)	0.54 (0.46-0.63)		0.87 (0.59-1.26)	0.91 (0.55-1.48)	
ApoB	6.55 (5.28-8.12)		5.01 (2.90-8.64)	3.97 (2.11-7.46)	4.66 (2.50-8.67)	7.22 (5.79-9.00)		2.74 (1.46-5.15)	2.81 (1.41-5.60)	2.38 (1.28-4.44)
Sqrt Lp(A)	1.09 (1.07-1.10)		1.03 (1.01-1.06)	1.04 (1.02-1.06)		1.11 (1.09-1.13)		1.04 (1.01-1.06)	1.04 (1.01-1.06)	1.04 (1.01-1.06)
Log CRP	1.44 (1.38-1.51)		1.17 (1.10-1.24)	1.11 (1.04-1.18)	1.10 (1.03-1.17)	1.47 (1.41-1.53)		1.22 (1.15-1.29)	1.16 (1.09-1.24)	1.18 (1.11-1.25)
Albumin	0.96 (0.95-0.97)		0.97 (0.95-0.99)	0.98 (0.96-1.01)	0.97 (0.96-0.99)	0.97 (0.95-0.98)		0.97 (0.94-0.99)	0.97 (0.95-1.00)	0.98 (0.96-1.00)
Log creatinine	1.49 (1.15-1.92)		1.09 (0.79-1.51)	1.23 (0.87-1.75)	1.21 (0.86-1.72)	1.86 (1.39-2.49)		1.18 (0.84-1.64)	1.24 (0.85-1.80)	
Uric acid	1.00 (1.00-1.00)		1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.01 (1.00-1.01)		1.00 (1.00-1.00)	1.00 (1.00-1.00)	
Glucose	1.10 (1.07-1.13)		1.04 (1.01-1.07)	0.97 (0.92-1.02)	0.96 (0.92-1.01)	1.16 (1.13-1.19)		1.04 (1.00-1.07)	0.96 (0.92-1.01)	0.97 (0.92-1.02)
HbA1c	1.50 (1.41-1.60)		1.26 (1.15-1.37)	1.34 (1.18-1.51)	1.34 (1.18-1.52)	1.59 (1.46-1.72)		1.23 (1.12-1.34)	1.30 (1.16-1.45)	1.27 (1.14-1.42)
ALP	1.00 (1.00-1.01)		1.00 (1.00-1.01)	1.00 (1.00-1.01)	1.00 (1.00-1.00)	1.01 (1.01-1.02)		1.00 (1.00-1.01)	1.00 (0.99-1.01)	

Table 2: Continued

	Males					Females				
	Univariable	TP	TP + biomarker	Full	Combined	Univariable	TP	TP + biomarker	Full	Combined
ALT	1.00 (1.00-1.01)		1.00 (0.99-1.00)	1.00 (0.99-1.01)		1.01 (1.01-1.01)		1.00 (1.00-1.00)	1.00 (0.99-1.01)	1.00 (0.99-1.00)
AST	1.00 (1.00-1.01)		1.00 (0.99-1.00)	1.00 (0.99-1.00)		1.01 (1.00-1.01)		1.00 (0.99-1.01)	1.00 (0.99-1.01)	
GGT	1.35 (1.26-1.44)		1.01 (0.93-1.10)	0.99 (0.89-1.11)		1.62 (1.50-1.74)		1.10 (0.99-1.22)	1.05 (0.92-1.20)	
Calcium	0.97 (0.68-1.38)		0.62 (0.41-0.94)	0.66 (0.40-1.10)		2.28 (1.65-3.17)		1.07 (0.71-1.63)	1.39 (0.84-2.28)	
Magnesium	1.00 (0.51-1.94)		0.67 (0.30-1.51)	0.88 (0.36-2.13)		0.90 (0.46-1.75)		0.55 (0.24-1.24)	0.77 (0.31-1.93)	0.82 (0.34-1.98)
Iron	0.99 (0.98-1.00)		0.99 (0.98-1.00)	1.00 (0.99-1.01)	0.99 (0.98-1.00)	0.99 (0.98-1.00)		0.99 (0.98-1.01)	1.00 (0.99-1.01)	
Transferrin	1.01 (1.00-1.01)		1.01 (1.00-1.02)	1.01 (1.00-1.02)		0.98 (0.97-0.99)		1.00 (0.99-1.01)	0.99 (0.98-1.01)	
Log ferritin	1.10 (1.04-1.16)		0.94 (0.88-1.00)	1.00 (0.92-1.08)		1.55 (1.47-1.64)		1.02 (0.95-1.09)	0.99 (0.91-1.08)	
Total bilirubin	0.96 (0.95-0.97)		1.00 (0.99-1.01)	1.00 (0.99-1.01)		0.95 (0.94-0.97)		1.00 (0.99-1.02)	1.01 (0.99-1.03)	

Values represent hazard ratios (95% Confidence interval). TP: traditional prediction model. SBP: systolic blood pressure, HDL: high-density lipoprotein, Apo: apolipoprotein, Lp(a): lipoprotein(a), CRP: C-reactive protein, HbA1c: glycated hemoglobin, ALP: alkaline phosphatase, ALT: alanine transaminase, AST: aspartate aminotransferase, GGT: gamma-glutamyl transferase.

Improvement in predictive performance

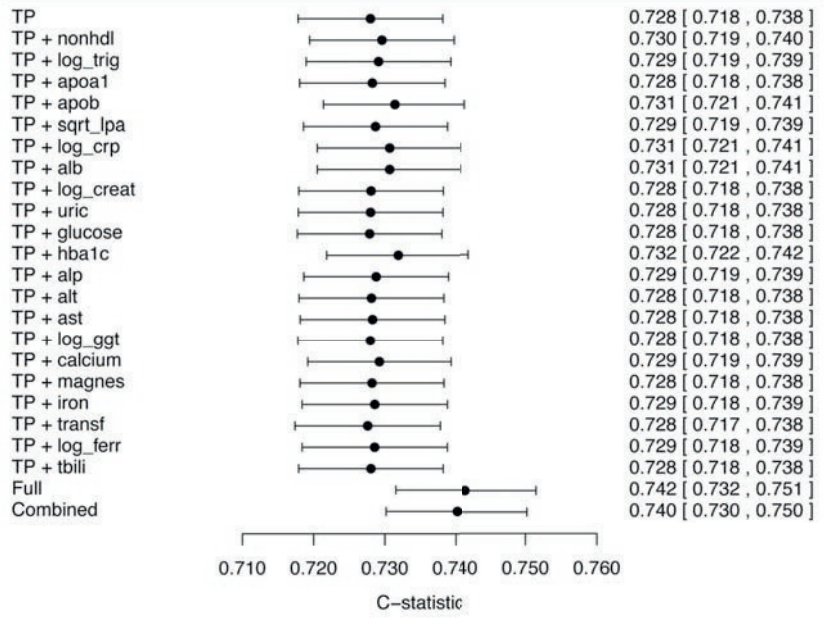
Figure 1 shows the c-statistic of the models including the traditional predictors plus each biomarker. Biomarkers with the highest increase in c-statistic compared to the traditional prediction model were apoB, CRP, and HbA1c for both males and females, as well as non-HDL cholesterol and albumin in males. The c-statistic of the traditional prediction model was lower in males compared to females (0.728 vs. 0.763, respectively), and the increase in c-statistic when adding a biomarker was overall larger for males.
In terms of calibration (Figure 2), apoB, albumin, glucose, and HbA1c resulted in an OE ratio closer to 1 for both males and females, as well as ALP, calcium, and iron for males and magnesium in females.

In terms of reclassification (Table 3) in people who experienced the event, apoB, CRP, and HbA1c improved reclassification in males, while CRP and glucose had most incremental value in females. The highest reclassification in non-cases was found for apoB (0.003), and creatinine (0.002) in males. No NRIs >0 were found for female non-cases. Other performance measures, such as R^2, BIC and Brier score showed consistent results with respect to the relative importance of the different biomarkers, but the differences between the traditional prediction model and the models with a biomarker added were small (Table S5).

Combined model

In males, the predictors non-HDL cholesterol, triglycerides, apoA1, apoB, CRP, albumin, creatinine, uric acid, glucose, HbA1c, ALP, and iron were in the top 5 of biomarkers with most improvement in either c-statistic, OE ratio, or NRI, while in females we selected apoB, Lp(a), CRP, albumin, glucose, HbA1c, ALT, and magnesium. Non-HDL cholesterol and apoB were highly correlated, and we chose to retain apoB because it displayed the largest incremental value on all performance statistics. Choosing other threshold values for risk categories for NRI calculation did not influence the choice of predictors (data not shown). In males, the c-statistic of the traditional versus the full versus the combined model was 0.728 vs. 0.742 vs. 0.740 (Figure 1) while the OE ratios were 1.010 vs. 0.993 vs. 0.995, respectively (Figure 2). In females the c-statistic was 0.763 vs. 0.772 vs. 0.771 (Figure 1) and OE ratio 1.116 vs. 1.115 vs. 1.116 (Figure 2) for the traditional, full, and combined models, respectively. The calibration plot showed no differences between the three models, both in males and females (Figure 3).

Males



Females

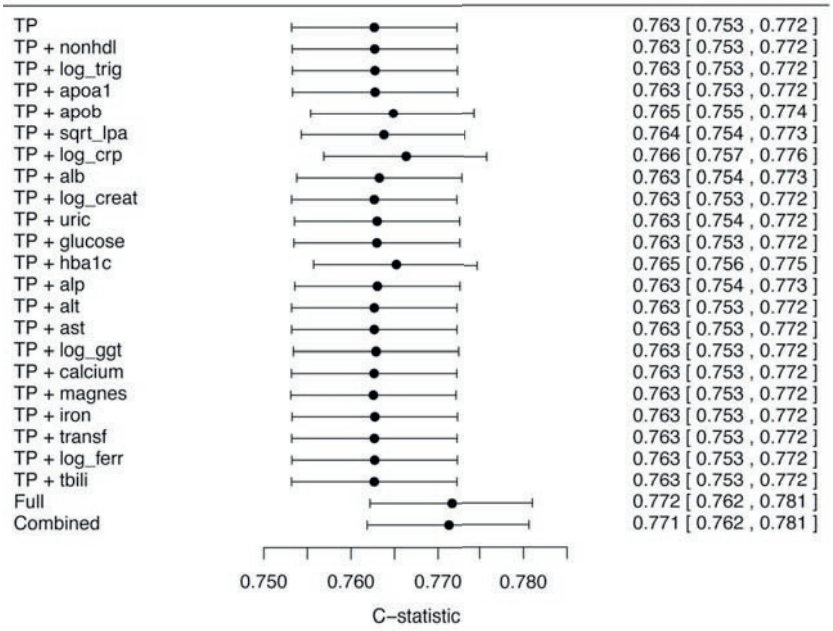
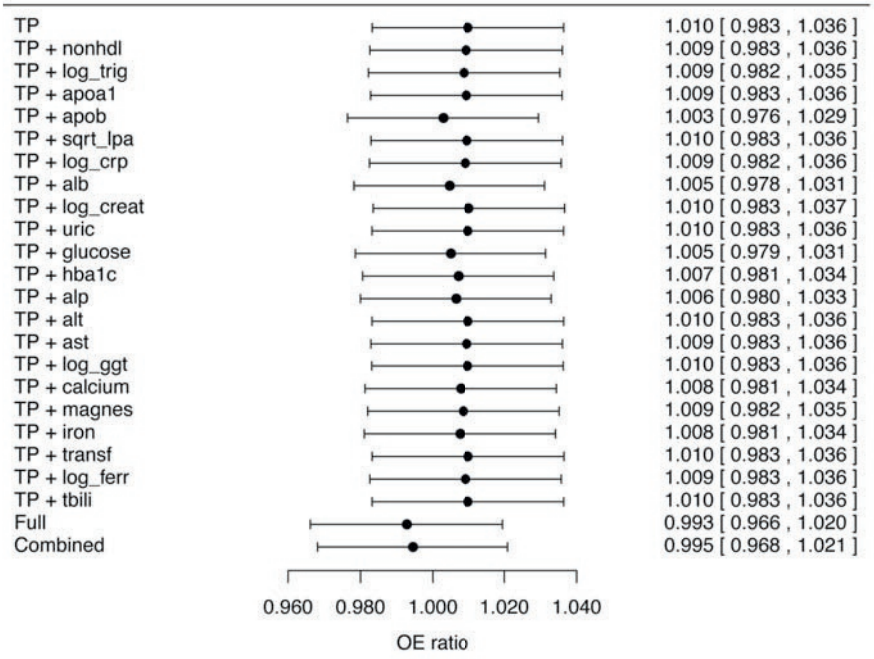


Figure 1: Forest plot with c-statistic of the traditional prediction model, traditional prediction models with one biomarker added, the combined model and the full model. TP (traditional prediction) model includes age, current smoking, diabetes, hypertension, log bmi, systolic blood pressure, log total cholesterol, log HDL cholesterol. Full model includes all predictors from the traditional prediction model plus all predictors listed in the table. Combined model includes all predictors from the traditional prediction model plus log triglycerides, apoB, log CRP, albumin, glucose, HbA1c, ALP, ALT, and iron in males, and apoB, sqrt Lp(a), log CRP, albumin, glucose, HbA1c, ALT, AST, and

magnesium in females. HDL: high-density lipoprotein, apo: apolipoprotein, Lp(a): lipoprotein(a), CRP: C-reactive protein, HbA1c: glycated hemoglobin, ALP: alkaline phosphatase, ALT: alanine transaminase, AST: aspartate aminotransferase, GGT: gamma-glutamyl transferase.

Males



Females

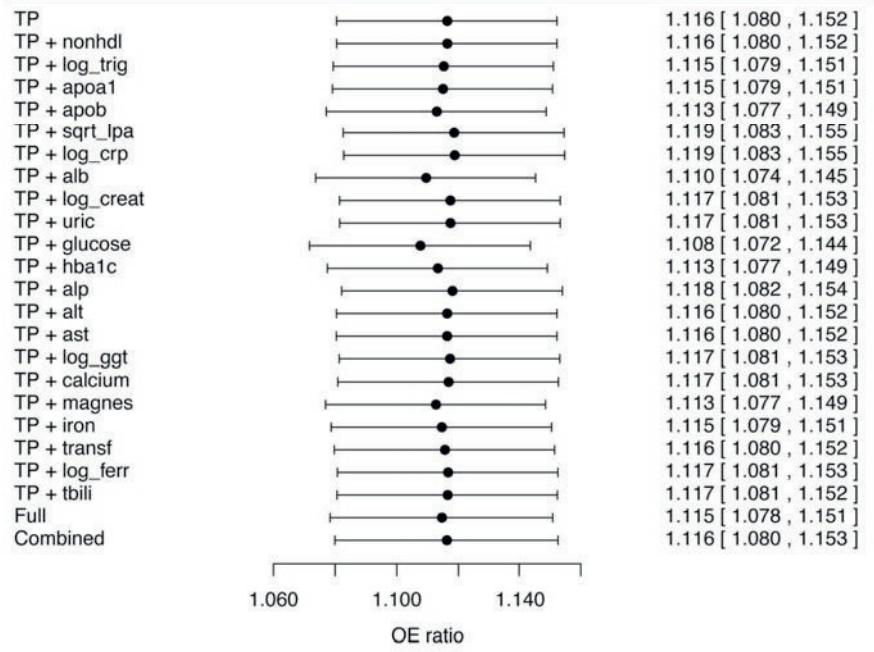


Figure 2: (previous page) Forest plot of the OE ratio of the traditional prediction model, traditional prediction models with one biomarker added, the combined model and the full model. TP (traditional prediction) model includes age, current smoking, diabetes, hypertension, log bmi, systolic blood pressure, log total cholesterol, log HDL cholesterol. Full model includes all predictors from the traditional prediction model plus all predictors listed in the table. Combined model includes all predictors from the traditional prediction model plus log triglycerides, apoB, log CRP, albumin, glucose, HbA1c, ALP, ALT, and iron in males, and apoB, sqrt Lp(a), log CRP, albumin, glucose, HbA1c, ALT, AST, and magnesium in females. HDL: high-density lipoprotein, apo: apolipoprotein, Lp(a): lipoprotein(a), CRP: C-reactive protein, HbA1c: glycated hemoglobin, ALP: alkaline phosphatase, ALT: alanine transaminase, AST: aspartate aminotransferase, GGT: gamma-glutamyl transferase.

Table 3: NRI, models with one biomarker added, combined model and full model, compared to model with traditional predictors only.

	Males		**Females**	
	Cases	**Non-cases**	**Cases**	**Non-cases**
TP	ref	ref	ref	ref
TP + Non-HDL cholesterol	-0.002	0.000	0.001	0.000
TP + Log triglycerides	0.005	0.001	0.002	0.000
TP + ApoA1	0.001	0.000	0.001	0.000
TP + ApoB	0.021	0.003	0.003	-0.001
TP + Sqrt Lp(A)	0.001	-0.002	0.000	-0.001
TP + Log CRP	0.016	-0.005	0.010	-0.002
TP + Albumin	0.005	-0.004	0.009	-0.002
TP + Log creatinine	-0.001	0.002	-0.001	0.000
TP + Uric acid	0.000	0.000	-0.003	0.000
TP + Glucose	0.008	0.000	0.013	0.000
TP + HbA1c	0.013	-0.003	0.007	0.000
TP + ALP	0.000	-0.001	-0.003	0.000
TP + ALT	0.002	0.000	0.000	0.000
TP + AST	0.001	0.000	0.000	0.000
TP + GGT	0.000	0.000	-0.004	-0.001
TP + Calcium	0.002	-0.004	0.000	0.000
TP + Magnesium	0.000	0.000	0.002	-0.001
TP + Iron	0.004	0.001	0.000	0.000
TP + Transferrin	-0.001	0.000	0.001	0.000
TP + Log ferritin	0.001	-0.001	0.000	0.000
TP + Total bilirubin	0.001	0.000	0.000	0.000
Full	0.058	-0.002	0.034	-0.003
Combined	0.048	-0.003	0.026	-0.001

TP (traditional prediction) model includes age, current smoking, diabetes, hypertension, log bmi, systolic blood pressure, log total cholesterol, log HDL cholesterol. Full model includes all predictors from the traditional prediction model plus all predictors listed in the table. Combined model includes all predictors from the traditional prediction model plus log triglycerides, apoB, log CRP, albumin, glucose, HbA1c, ALP, ALT, and iron in males, and apoB, sqrt Lp(a), log CRP, albumin, glucose, HbA1c, ALT, AST, and magnesium in females. HDL: high-density lipoprotein, apo: apolipoprotein, Lp(a): lipoprotein(a), CRP: C-reactive protein, HbA1c: glycated hemoglobin, ALP: alkaline phosphatase, ALT: alanine transaminase, AST: aspartate aminotransferase, GGT: gamma-glutamyl transferase.

Males ***Females***

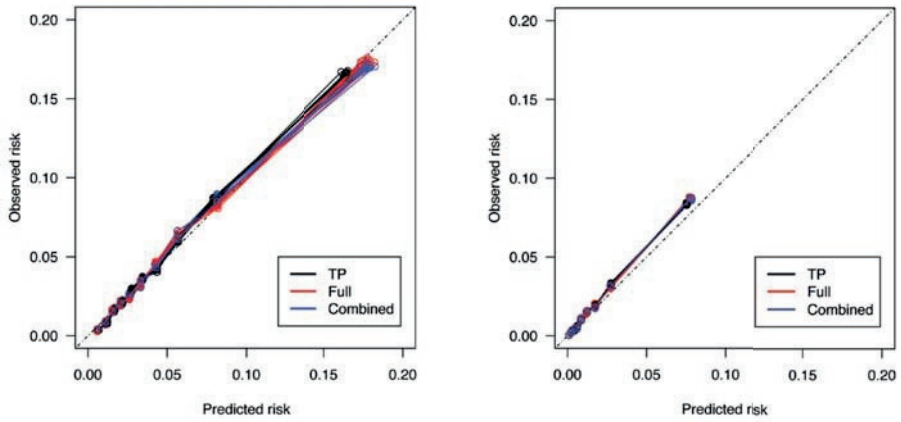


Figure 3: calibration plots of the traditional prediction model, full model and combined model. Every line represents one imputation. TP: traditional prediction model.

Differences between countries

Apparent performance of the models stratified by country revealed that different predictors have most incremental value in different countries (Table S6). For example, apoB, CRP, and calcium had most incremental value in terms of discrimination in males in Italy, while in the Netherlands Lp(a), glucose and HbA1c had most incremental value. C-statistics of the combined model were lowest in the UK, the Netherlands, and Oxford (Figure S1). OE ratios were below 1 (meaning overestimation of risks) in the UK, Netherlands (females), Greece, Denmark and Sweden, and above 1 in Italy (males), Spain, Netherlands (males), and Oxford (Figure S2).

Discussion

Summary of findings

The aim of this study was to identify biomarkers that have incremental value in predicting the risk of CHD within 10 years on top of traditional predictors. In males, the combination of non-HDL cholesterol, triglycerides, apoA1, apoB, CRP, albumin, creatinine, uric acid, glucose, HbA1c, ALP, and iron had the most incremental value, while these were in females apoB, LP(a), CRP, albumin, glucose, HbA1c, ALT, and magnesium. Improvement in predictive performance measures and reclassification of subjects that might potentially lead to different treatment decisions was, however, limited.

Comparison with literature

Multiple primary studies and systematic reviews have been published, reporting on the incremental value of various biomarkers on the prediction of CHD or CVD. In agreement with our results, these studies report incremental value of LP(a), HbA1c, ALP and ALT, iron, CRP, and albumin.[16,19-21,24,25] We could not confirm incremental value of GGT and ferritin.[20,22]

Strengths and limitations

In contrast to these aforementioned studies, we investigated the potential incremental value of a large set of biomarkers in a single study population, allowing for a direct comparison of their relative importance. Other main strengths of this study are the large sample size, and the availability of data from various countries from Europe, allowing for differences across countries at different risk of CVD, and thus good generalizability of our results.

We also acknowledge some limitations. Firstly, although data collection was standardized within countries, there were differences in measurement of traditional predictors and outcome between countries. This was accounted for by using a stratified baseline hazard per country. Systolic blood pressure was not registered in some of the Spanish cohorts at all, which we accounted for using multiple imputation.

Secondly, although the EPIC cohort is representative for the European population,[29] blood samples were available for only part of the patients from the EPIC cohort which may lead to selection bias. Also, within the EPIC-CVD cohort healthy people may be slightly overrepresented.[41]

Thirdly, we assessed incremental value for prediction of CHD risk and not CVD, whereas differences in predictors between these two outcomes have been reported.[9,42] Therefore we cannot preclude that some of the biomarkers that showed incremental value in prediction of CHD may not improve CVD prediction models. However, given that the biomarkers we identified had only limited incremental value for CHD prediction, it is not to be expected that they would drastically improve prediction of CVD.

Fourthly, the EPIC-CVD cohort included participants between 1991 and 1999. Changes have occurred in treatment strategies, e.g. over time more people are being treated with risk lowering drugs.[43] Information regarding treatment use at baseline was frequently missing in our study, and treatment use during follow-up was not collected. Ignoring the effects of treatment may typically lead to underestimation of risk in high-risk patients[44-46] and we cannot exclude that this has influenced relative importance of certain biomarkers. For example, if high risk individuals have higher CRP values, and they had a higher chance of receiving risk-lowering treatment during the EPIC study follow-up, the association between CRP and CHD might have been attenuated in this study. However, it should be noted that our aim was not to develop a new prediction model, but rather to directly compare models with and without biomarkers within the same dataset.

Finally, in our investigation we did not take into account direct added clinical benefit and improved cost-effectiveness of the biomarkers. The performance measures we used to decide on incremental value, such as the c-statistic and NRI have been criticized in this regard. Improvement in the c-statistic might be small and therefore predictors with incremental value might be missed.[13,14,47,48] As the NRI is highly dependent on which risk thresholds are chosen,[47] we calculated the NRI for different risk categories, which did not influence the choice of predictors. Moreover, the other performance measures we investigated, such as R^2, BIC and Brier score, showed similar results with respect to the relative importance of the different biomarkers. The absolute differences between the traditional model and the extended models were small, suggesting limited improvement in overall fit and overall predictive ability. Performance measures used in development of prediction models which focus more on the clinical benefit, such as decision curve analysis,[49] are currently not available for studies with a case-cohort design.

Clinical implications

We identified a large number of predictors with potential incremental value in predicting CHD. Although we did not perform a formal impact study,[50-52] considering the large number of biomarkers that was needed to demonstrate a limited improvement in model discrimination, calibration and reclassification, it is unlikely that it will be cost-effective to routinely measure all these predictors on a population level. Such impact modelling, addressing the extent to which the biomarkers with added value indeed change therapeutic decision making and subsequent individual outcomes, is the next step to determine which biomarker should be measured in which subgroups. A different approach might be to measure the additional predictors with added value in the group of patients with intermediate risk based on the traditional predictors, as has been addressed for other tests in this field.[52-54]

Also, we observed large differences in predictive performance of traditional predictors and in incremental value of predictors between countries, which was not fully accounted for by a stratified baseline hazard. Hence, given these large differences

between countries, a single prediction model for all European countries might be not feasible.

Implications for further research

Previous studies have shown that the performance of current prediction models is often poor.[10,55,56] Since there is already an overabundance of prediction models for cardiovascular disease, we have previously advised to focus on improving these currently available prediction models, and not develop new models from scratch.[6] One way to do this is to add new predictors to current models. An alternative strategy, which might be much more effective, could be to tailor existing prediction models based on traditional predictors only to different countries or local settings using model updating strategies, such as recalibration of the model intercept or baseline hazard.[57-59] Additionally, when the performance of these models is sufficient (e.g. no overestimation of predicted risks), the models can be used to select a group of patients at intermediate risk, for which it is unclear whether these should be treated or not. Future research should focus on incremental value of predictors in this selective intermediate risk group for which traditional predictors did not result in clear treatment decisions.[52,53]

Conclusion

We found that many biomarkers need to be measured to gain limited increase in predictive performance over the traditional prediction model for the prediction of 10-year risk of CHD. Based on our current findings we would not advise to add certain biomarkers to traditional prediction models. Additional impact analyses, including cost-effectiveness studies, are needed to determine whether this improvement in predictive performance as compared to using only the traditional predictors, improves therapeutic decision making and subsequent patient outcome.

References

1. Wilkins E WL, Wickramasinghe K, Bhatnagar P, Leal J, Luengo-Fernandez R, Burns R, Rayner M, Townsend N European Cardiovascular Disease Statistics 2017. Brussels: European Heart Network, 2017.
2. Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014;129(25 Suppl 2):S49-73.
3. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* 2002;106(25):3143-421.
4. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J* 2003;24(11):987-1003.
5. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;357:j2099.
6. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;353:i2416.
7. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. *Am Heart J* 1991;121(1 Pt 2):293-8.
8. D'Agostino RB, Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008;117(6):743-53.
9. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97(18):1837-47.
10. Brindle P, Beswick A, Fahey T, Ebrahim S. Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. *Heart* 2006;92(12):1752-9.
11. Cook NR, Ridker PM. Calibration of the Pooled Cohort Equations for Atherosclerotic Cardiovascular Disease: An Update. *Ann Intern Med* 2016.
12. Polonsky TS, Greenland P. CVD screening in low-risk, asymptomatic adults: clinical trials needed. *Nat Rev Cardiol* 2012;9(10):599-604.
13. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Invest* 2012;42(2):216-28.

14. Austin PC, Pencinca MJ, Steyerberg EW. Predictive accuracy of novel risk factors and markers: A simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model. *Stat Methods Med Res* 2017;26(3):1053-77.
15. Di Angelantonio E, Sarwar N, Perry P, Kaptoge S, Ray KK, Thompson A, et al. Major lipids, apolipoproteins, and risk of vascular disease. *JAMA* 2009;302(18):1993-2000.
16. Erqou S, Kaptoge S, Perry PL, Di Angelantonio E, Thompson A, White IR, et al. Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. *JAMA* 2009;302(4):412-23.
17. Erqou S, Thompson A, Di Angelantonio E, Saleheen D, Kaptoge S, Marcovina S, et al. Apolipoprotein(a) isoforms and the risk of vascular disease: systematic review of 40 studies involving 58,000 participants. *J Am Coll Cardiol* 2010;55(19):2160-7.
18. Thompson A, Danesh J. Associations between apolipoprotein B, apolipoprotein AI, the apolipoprotein B/AI ratio and coronary heart disease: a literature-based meta-analysis of prospective studies. *J Intern Med* 2006;259(5):481-92.
19. Sarwar N, Aspelund T, Eiriksdottir G, Gobin R, Seshasai SR, Forouhi NG, et al. Markers of dysglycaemia and risk of coronary heart disease in people without diabetes: Reykjavik prospective study and systematic review. *PLoS Med* 2010;7(5):e1000278.
20. Kunutsor SK, Apekey TA, Khan H. Liver enzymes and risk of cardiovascular disease in the general population: a meta-analysis of prospective cohort studies. *Atherosclerosis* 2014;236(1):7-17.
21. Lapice E, Masulli M, Vaccaro O. Iron deficiency and cardiovascular disease: an updated review of the evidence. *Curr Atheroscler Rep* 2013;15(10):358.
22. Zhou Y, Liu T, Tian C, Kang P, Jia C. Association of serum ferritin with coronary artery disease. *Clin Biochem* 2012;45(16-17):1336-41.
23. Li X, Meng X, Timofeeva M, Tzoulaki I, Tsilidis KK, Ioannidis PA, et al. Serum uric acid levels and multiple health outcomes: umbrella review of evidence from observational studies, randomised controlled trials, and Mendelian randomisation studies. *BMJ* 2017;357:j2376.
24. Kaptoge S, Di Angelantonio E, Lowe G, Pepys MB, Thompson SG, Collins R, et al. C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis. *Lancet* 2010;375(9709):132-40.
25. Danesh J, Collins R, Appleby P, Peto R. Association of fibrinogen, C-reactive protein, albumin, or leukocyte count with coronary heart disease: meta-analyses of prospective studies. *JAMA* 1998;279(18):1477-82.
26. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162(1):55-63.

27. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.
28. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* 2002;5(6b):1113-24.
29. Danesh J, Saracci R, Berglund G, Feskens E, Overvad K, Panico S, et al. EPIC-Heart: the cardiovascular component of a prospective study of nutritional, lifestyle and biological factors in 520,000 middle-aged participants from 10 European countries. *Eur J Epidemiol* 2007;22(2):129-41.
30. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986;73(1):1-11.
31. Harrell F. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*: Springer, 2015.
32. Xue X, Xie X, Gunter M, Rohan TE, Wassertheil-Smoller S, Ho GY, et al. Testing the proportional hazards assumption in case-cohort analysis. *BMC Med Res Methodol* 2013;13:88.
33. Sanderson J, Thompson SG, White IR, Aspelund T, Pennells L. Derivation and assessment of risk prediction models using case-cohort data. *BMC Med Res Methodol* 2013;13:113.
34. Steyerberg E. *Clinical prediction models: a practical approach to development, validation, and updating*: Springer Science & Business Media, 2008.
35. Tzoulaki I, Liberopoulos G, Ioannidis JP. Use of reclassification for assessment of improved prediction: an empirical evaluation. *Int J Epidemiol* 2011;40(4):1094-105.
36. Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *J Stat Softw* 2011;45(3).
37. White IR, Royston P. Imputing missing covariate values for the Cox model. *Stat Med* 2009;28(15):1982-98.
38. Rubin DB. *Multiple imputation for nonresponse in surveys*: John Wiley & Sons, 2004.
39. Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res* 2017:962280217705678.
40. R: A language and environment for statistical computing [program]. Vienna, Austria: R Foundation for Statistical Computing, 2016.
41. Lassale C, Tzoulaki I, Moons KGM, Sweeting M, Boer J, Johnson L, et al. Separate and combined associations of obesity and metabolic health with coronary heart disease: a pan-European case-cohort analysis. *Eur Heart J* 2017.

42. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham Study. *Stroke* 1991;22(3):312-8.
43. Bhatnagar P, Wickramasinghe K, Wilkins E, Townsend N. Trends in the epidemiology of cardiovascular disease in the UK. *Heart* 2016;102(24):1945-52.
44. Groenwold RH, Moons KG, Pajouheshnia R, Altman DG, Collins GS, Debray TP, et al. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *J Clin Epidemiol* 2016;78:90-100.
45. Pajouheshnia R, Peelen LM, Moons KGM, Reitsma JB, Groenwold RHH. Accounting for treatment use when validating a prognostic model: a simulation study. *BMC Med Res Methodol* 2017;17(1):103.
46. Peek N, Sperrin M, Mamas M, Van Staa T, Buchan I. Hari Seldon, QRISK3, and the prediction paradox. *BMJ* 2017;357:j2099.
47. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27(2):157-72; discussion 207-12.
48. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. *Stat Med* 2013;32(9):1467-82.
49. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016;352:i6.
50. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009;338:b606.
51. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;98(9):691-8.
52. van Giessen A, Moons KG, de Wit GA, Verschuren WM, Boer JM, Koffijberg H. Tailoring the implementation of new biomarkers based on their added predictive value in subgroups of individuals. *PLoS One* 2015;10(1):e0114020.
53. Paynter NP, Cook NR. Adding tests to risk based guidelines: evaluating improvements in prediction for an intermediate risk group. *BMJ* 2016;354:i4450.
54. Den Ruijter HM, Peters SA, Anderson TJ, Britton AR, Dekker JM, Eijkemans MJ, et al. Common carotid intima-media thickness measurements in cardiovascular risk prediction: a meta-analysis. *JAMA* 2012;308(8):796-803.
55. Cook NR, Ridker PM. Further insight into the cardiovascular risk calculator: the roles of statins, revascularizations, and underascertainment in the Women's Health Study. *JAMA* 2014;174(12):1964-71.
56. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol* 2015;68(1):25-34.

57. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;61(1):76-86.
58. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol* 2008;61(11):1085-94.
59. Nieboer D, Vergouwe Y, Ankerst DP, Roobol MJ, Steyerberg EW. Improving prediction models with new markers: a comparison of updating strategies. *BMC Med Res Methodol* 2016;16(1):128.

Supplemental material

Table S1: TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist Item	Page*
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	3
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	3
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	4
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	4, TS2
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	4, TS2
	5b	D;V	Describe eligibility criteria for participants.	4, TS2
	5c	D;V	Give details of treatments received, if relevant.	T 1, TS3
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	5, T S2
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	5
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	4, 5
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	4, 5
Sample size	8	D;V	Explain how the study size was arrived at.	NA
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	7, 8

Table S1: Continued

Section/Topic	Item		Checklist Item	Page*
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	6
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	6, 7
	10c	V	For validation, describe how the predictions were calculated.	NA
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	6, 7
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	7
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	NA
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	9
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	9, T 1
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	T 1
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	T 2
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	NA
	15b	D	Explain how to the use the prediction model.	NA
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	14-19
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	NA

Table S1: Continued

Section/Topic	Item		Checklist Item	Page*
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	20, 21
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	NA
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	20-22
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	21,22
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	27
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	1

*Values correspond to page numbers in original article. T: table, NA: not applicable

Table S2: Description of centres

Country	Centre	N subcohort	N cases added	Recruitment period (years)	End of FU		Eligibility criteria
					Nonfatal	*Fatal*	
Denmark	Aarhus	534	449	1995-1997	Dec-09	Mar-10	Men and women aged 50-64, without prevalent cancer
	Copenhagen	1200	1130	1993-1997	Dec-09	Mar-10	Men and women aged 50-64, without prevalent cancer
Germany	Heidelberg	866	337	1994-1998	May-10	May-10	Men aged 40-65, women aged 35-65

Table S2: Continued

Country	Centre	N subcohort	N cases added	Recruitment period (years)	End of FU		Eligibility criteria
					Nonfatal	*Fatal*	
	Potsdam	1135	268	1994-1998	Nov-08	Nov-08	Men aged 40-65, women aged 35-65
Greece	Greece	1159	318	1994-1999	Dec-09	Dec-09	Apparently healthy men and women aged 25-82
Italy	Florence	533	152	1993-1998	Dec-03	Dec-03	Breast cancer screening participants and general population; men aged 35-64, women aged 35-64, without prevalent cancer
	Ragusa	329	179	1993-1997	Dec-07	Dec-09	Blood donors and general population; men aged 40-65, women aged 35-65
	Turin	536	184	1993-1998	Dec-09	Dec-09	Blood donors and general population; men aged 40-74, women aged 35-74, without prevalent cancer
	Naples	217	87	1993-1997	Dec-06	Dec-06	Women aged 30-69
	Varese	360	252	1993-1997	Dec-06	Dec-06	Men aged 40-65, women aged 35-65

Table S2: Continued

Country	Centre	N subcohort	N cases added	Recruitment period (years)	End of FU		Eligibility criteria
					Nonfatal	*Fatal*	
Netherlands	Bilthoven	376	635	1993-1997	Dec-07	Dec-07	Men and women aged 20-65
	Utrecht	872	812	1993-1997	Dec-07	Dec-07	Population-based breast cancer screening participants aged 49-70
Spain	Asturias	773	255	1992–1995	Dec-06	Dec-06	Blood donors and general population; men aged 40-64, women aged 35-64
	Granada	535	158	1993-1996	Dec-08	Dec-08	Blood donors and general population; men aged 40-64, women aged 35-64
	Murcia	765	143	1992-1996	Dec-08	Dec-08	Blood donors and their parents, and general population; men aged 40-65, women aged 35-65
	Navarra	772	274	1992-1995	Dec-08	Dec-08	Blood donors and general population; men aged 40-65, women aged 35-65

Table S2: Continued

Country	Centre	N subcohort	N cases added	Recruitment period (years)	End of FU		Eligibility criteria
					Nonfatal	*Fatal*	
	San Sebastian	769	284	1992-1995	Dec-08	Dec-08	Blood donors and employees of selected enterprises; men aged 40-65, women aged 35-65
Sweden	Malmo	1667	1493	1991-1996	Dec-08	Dec-08	Men aged 50-72, women aged 46-72
	Umea	951	587	1992-1996	Dec-06	Dec-06	Men and women aged 30, 40, 50 or 60
UK	Cambridge	817	1550	1993-1998	Dec-06	Dec-06	Men and women aged 45-74
	Oxford	298	1009	1994-1997	Dec-09	Dec-10	Vegetarians, vegans and other health-conscious individuals aged 20+ and general population aged 40-65
France	France	569	38	1993–1997	Dec-09	Dec-10	Women aged 40-65
Total		16242	10594	1991-1999			

FU: follow-up, UK: United Kingdom.

Table S3: baseline characteristics per country

France				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
N	0	0	510	36
Age (years)			56.3 (6.6)	62.4 (6.0)
Current smoker			48 (9.4%)	4 (11.1%)
BMI (kg/m2)			23.0 (3.6)	22.5 (3.8)
Diabetes			5 (1.0%)	1 (2.8%)
Hypertension			145 (28.4%)	21 (58.3%)
SBP (mmHg)			124.6 (17.8)	136.5 (20.0)
Use of antihypertensive medication at baseline			27 (5.3%)	5 (13.9%)
Use of lipid lowering medication at baseline			31 (6.1%)	10 (27.8%)
Total cholesterol (mmol/l)			6.0 (0.9)	6.6 (1.0)
HDL cholesterol (mmol/l)			1.8 (0.4)	1.8 (0.5)
non-HDL cholesterol (mmol/l)			4.2 (0.9)	4.8 (1.1)
Triglycerides (mmol/l)			0.8 (0.6-1.1)	0.9 (0.7-1.2)
ApoA1 (g/l)			1.7 (0.3)	1.7 (0.2)
ApoB (g/l)			0.9 (0.2)	1.1 (0.2)
Lp(a) (mg/dl)			29.2 (16.0-57.7)	42.8 (21.4-68.5)
CRP (mg/l)			0.7 (0.4-1.7)	1.0 (0.4-4.8)
Albumin (g/l)			45.6 (2.6)	45.8 (3.3)
Creatinine (umol/l)			65.0 (59.0-72.0)	67.5 (60.8-74.8)
Uric acid (umol/l)			248.9 (58.1)	291.1 (73.5)
Glucose (mmol/l)			4.0 (3.6-4.4)	4.0 (3.6-4.2)
HbA1c (%)			5.4 (5.3-5.7)	5.6 (5.4-5.8)
ALP(iU/l)			54.0 (45.0-67.0)	60.0 (46.8-68.2)
ALT (iU/l)			16.0 (13.0-20.0)	17.0 (13.0-24.0)
AST (iU/l)			26.0 (24.0-31.0)	30.0 (26.0-33.2)
GGT (iU/l)			17.0 (14.0-24.0)	20.5 (15.8-25.8)
Calcium (mmol/l)			2.4 (0.1)	2.4 (0.1)
Magnesium (mmol/l)			0.9 (0.1)	0.9 (0.1)

Table S3: Continued

France				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
Iron (umol/l)			18.3 (5.8)	18.1 (5.4)
Transferrin (umol/l)			33.9 (5.2)	34.4 (6.3)
Ferritin (pmol/l)			191.0 (98.9-301.1)	255.1 (154.0-388.7)
Total bilirubin (umol/l)			7.0 (6.0-9.0)	7.0 (6.0-9.2)

Italy				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
N	573	432	1150	354
Age (years)	49.7 (7.5)	53.8 (7.2)	50.4 (8.1)	55.9 (7.4)
Current smoker	176 (30.7%)	186 (43.1%)	288 (25.0%)	119 (33.6%)
BMI (kg/m2)	26.2 (3.5)	27.4 (3.4)	25.6 (4.3)	27.3 (4.7)
Diabetes	16 (2.8%)	26 (6.0%)	17 (1.5%)	32 (9.0%)
Hypertension	202 (35.3%)	229 (53.0%)	366 (31.8%)	228 (64.4%)
SBP (mmHg)	131.7 (16.3)	138.2 (17.7)	128.0 (18.3)	141.1 (22.1)
Use of antihypertensive medication at baseline	63 (11.0%)	100 (23.1%)	141 (12.3%)	124 (35.0%)
Use of lipid lowering medication at baseline	12 (2.1%)	27 (6.2%)	28 (2.4%)	28 (7.9%)
Total cholesterol (mmol/l)	5.8 (1.0)	6.1 (1.1)	6.0 (1.1)	6.5 (1.3)
HDL cholesterol (mmol/l)	1.3 (0.3)	1.1 (0.3)	1.6 (0.4)	1.4 (0.4)
non-HDL cholesterol (mmol/l)	4.5 (1.0)	5.0 (1.1)	4.4 (1.1)	5.1 (1.4)
Triglycerides (mmol/l)	1.1 (0.8-1.6)	1.4 (1.0-2.0)	0.9 (0.7-1.2)	1.1 (0.9-1.6)
ApoA1 (g/l)	1.4 (0.2)	1.3 (0.2)	1.6 (0.3)	1.6 (0.3)
ApoB (g/l)	1.0 (0.2)	1.2 (0.3)	1.0 (0.2)	1.2 (0.3)
Lp(a) (mg/dl)	43.2 (22.5-70.6)	49.8 (26.9-78.5)	37.2 (18.9-62.1)	47.7 (26.0-91.9)
CRP (mg/l)	1.0 (0.6-1.9)	1.6 (0.8-3.4)	1.1 (0.5-2.5)	1.9 (0.9-3.8)
Albumin (g/l)	47.1 (2.9)	46.2 (2.9)	46.5 (3.5)	46.2 (3.8)

Table S3: Continued

Italy				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
Creatinine (umol/l)	77.0 (71.0-85.0)	78.0 (69.0-86.0)	61.0 (55.0-68.0)	60.0 (54.0-68.0)
Uric acid (umol/l)	328.3 (69.1)	340.3 (73.7)	243.2 (61.6)	260.9 (68.2)
Glucose (mmol/l)	5.3 (4.9-5.7)	5.3 (4.9-6.0)	5.0 (4.7-5.4)	5.1 (4.7-5.7)
HbA1c (%)	5.4 (5.3-5.7)	5.6 (5.3-5.9)	5.4 (5.3-5.7)	5.7 (5.4-6.1)
ALP(iU/l)	63.0 (54.0-75.0)	67.0 (58.0-79.0)	62.0 (50.0-77.0)	75.0 (62.2-89.0)
ALT (iU/l)	22.0 (18.0-31.0)	23.0 (18.0-30.0)	17.0 (13.0-22.0)	18.0 (14.0-24.0)
AST (iU/l)	28.0 (24.0-32.0)	28.0 (24.0-32.0)	25.0 (22.0-29.0)	26.0 (23.0-30.0)
GGT (iU/l)	27.0 (20.0-40.0)	30.0 (22.0-44.0)	16.0 (12.0-22.0)	19.0 (14.2-27.0)
Calcium (mmol/l)	2.5 (0.1)	2.4 (0.1)	2.5 (0.2)	2.4 (0.1)
Magnesium (mmol/l)	0.8 (0.1)	0.9 (0.1)	0.8 (0.1)	0.8 (0.1)
Iron (umol/l)	18.4 (5.7)	17.8 (5.6)	16.6 (6.6)	17.1 (6.1)
Transferrin (umol/l)	34.7 (4.7)	35.0 (5.2)	36.3 (6.0)	36.0 (5.7)
Ferritin (pmol/l)	256.2 (134.8-521.3)	266.3 (146.1-510.7)	125.8 (58.4-224.7)	203.3 (103.4-323.0)
Total bilirubin (umol/l)	9.0 (7.0-13.0)	8.0 (6.0-11.0)	7.0 (6.0-10.0)	7.0 (6.0-9.0)

Table S3: Continued

Spain				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
N	154	90	354	55
Age (years)	52.2 (7.0)	54.8 (6.9)	49.5 (8.7)	57.1 (7.8)
Current smoker	61 (39.6%)	45 (50.0%)	58 (16.4%)	9 (16.4%)
BMI (kg/m2)	27.6 (3.2)	28.6 (3.5)	28.3 (4.8)	29.5 (4.5)
Diabetes	19 (12.3%)	17 (18.9%)	20 (5.6%)	11 (20.0%)
Hypertension	64 (41.6%)	61 (67.8%)	111 (31.4%)	43 (78.2%)
SBP (mmHg)	132.1 (17.4)	140.6 (23.0)	124.3 (18.3)	138.8 (21.8)
Use of antihypertensive medication at baseline	12 (7.8%)	34 (37.8%)	51 (14.4%)	35 (63.6%)
Use of lipid lowering medication at baseline	3 (1.9%)	5 (5.6%)	5 (1.4%)	8 (14.5%)
Total cholesterol (mmol/l)	6.0 (1.1)	6.2 (1.2)	5.8 (1.1)	6.2 (1.0)
HDL cholesterol (mmol/l)	1.3 (0.3)	1.2 (0.3)	1.6 (0.4)	1.5 (0.4)
non-HDL cholesterol (mmol/l)	4.6 (1.1)	5.0 (1.2)	4.3 (1.1)	4.8 (1.0)
Triglycerides (mmol/l)	1.2 (0.9-1.8)	1.4 (0.9-2.3)	0.9 (0.7-1.4)	1.2 (0.9-1.5)
ApoA1 (g/l)	1.5 (0.2)	1.4 (0.2)	1.6 (0.2)	1.5 (0.2)
ApoB (g/l)	1.1 (0.2)	1.2 (0.3)	1.0 (0.2)	1.1 (0.2)
Lp(a) (mg/dl)	48.6 (26.1-74.2)	55.8 (35.8-81.3)	45.5 (26.5-77.2)	38.8 (24.2-92.4)
CRP (mg/l)	1.0 (0.6-1.7)	1.5 (0.8-3.5)	1.3 (0.7-2.6)	2.4 (0.9-5.6)
Albumin (g/l)	48.1 (2.8)	47.9 (2.7)	46.6 (2.7)	46.3 (3.2)
Creatinine (umol/l)	76.0 (69.0-85.8)	76.5 (67.2-85.8)	60.0 (53.0-68.0)	62.0 (54.5-68.0)
Uric acid (umol/l)	342.3 (76.2)	359.1 (75.0)	258.8 (68.6)	286.5 (69.9)
Glucose (mmol/l)	5.2 (4.7-5.7)	5.1 (4.5-6.1)	4.7 (4.2-5.2)	4.8 (4.2-5.4)
HbA1c (%)	5.4 (5.2-5.6)	5.5 (5.3-5.9)	5.4 (5.2-5.6)	5.7 (5.4-6.0)
ALP(iU/l)	61.5 (52.0-70.0)	68.5 (62.0-77.8)	63.0 (51.0-79.0)	68.0 (59.5-81.5)
ALT (iU/l)	23.5 (18.0-31.0)	25.5 (20.0-35.0)	17.0 (13.0-21.8)	21.0 (18.0-27.0)
AST (iU/l)	29.5 (26.0-36.0)	32.0 (27.0-37.8)	25.0 (22.0-29.8)	27.0 (24.0-32.0)

Table S3: Continued

Spain				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
GGT (iU/l)	27.0 (19.0-39.8)	32.0 (25.0-45.8)	16.0 (13.0-21.0)	19.0 (14.0-28.5)
Calcium (mmol/l)	2.5 (0.1)	2.5 (0.1)	2.5 (0.1)	2.5 (0.2)
Magnesium (mmol/l)	0.9 (0.1)	0.9 (0.1)	0.9 (0.1)	0.9 (0.1)
Iron (umol/l)	18.4 (5.3)	19.2 (5.9)	15.2 (5.4)	14.8 (5.8)
Transferrin (umol/l)	34.6 (4.2)	34.3 (4.3)	35.9 (5.7)	36.4 (6.0)
Ferritin (pmol/l)	392.1 (216.3-688.8)	450.5 (301.7-677.0)	111.2 (49.4-202.2)	157.3 (82.0-296.6)
Total bilirubin (umol/l)	8.0 (6.0-11.0)	8.0 (6.0-10.8)	6.5 (5.0-9.0)	6.0 (4.0-8.0)

United Kingdom				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
N	306	1006	455	562
Age (years)	57.8 (8.7)	62.7 (7.7)	56.7 (8.8)	63.8 (7.2)
Current smoker	58 (19.0%)	231 (23.0%)	63 (13.8%)	94 (16.7%)
BMI (kg/m2)	25.7 (3.2)	26.7 (3.4)	25.4 (4.1)	26.7 (4.6)
Diabetes	9 (2.9%)	69 (6.9%)	5 (1.1%)	28 (5.0%)
Hypertension	127 (41.5%)	650 (64.6%)	171 (37.6%)	348 (61.9%)
SBP (mmHg)	134.7 (17.1)	143.4 (18.5)	131.5 (18.7)	141.7 (20.0)
Use of antihypertensive medication at baseline	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Use of lipid lowering medication at baseline	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Total cholesterol (mmol/l)	6.0 (1.1)	6.3 (1.1)	6.2 (1.2)	6.7 (1.2)
HDL cholesterol (mmol/l)	1.3 (0.3)	1.2 (0.3)	1.6 (0.4)	1.4 (0.4)
non-HDL cholesterol (mmol/l)	4.7 (1.1)	5.1 (1.1)	4.6 (1.2)	5.3 (1.3)
Triglycerides (mmol/l)	1.6 (1.1-2.2)	1.7 (1.2-2.4)	1.2 (0.8-1.8)	1.6 (1.2-2.3)
ApoA1 (g/l)	1.4 (0.2)	1.4 (0.2)	1.6 (0.3)	1.6 (0.3)
ApoB (g/l)	1.0 (0.2)	1.1 (0.2)	1.0 (0.3)	1.2 (0.3)

Table S3: Continued

United Kingdom

	Males		Females	
	No CHD	**CHD**	**No CHD**	**CHD**
Lp(a) (mg/dl)	47.5 (27.2-78.9)	55.4 (33.4-92.0)	42.4 (24.5-76.2)	58.4 (32.3-108.6)
CRP (mg/l)	1.0 (0.5-1.8)	1.7 (0.8-3.7)	1.0 (0.5-2.4)	1.9 (0.8-4.0)
Albumin (g/l)	46.5 (2.8)	46.3 (2.8)	46.0 (2.9)	45.8 (2.7)
Creatinine (umol/l)	81.0 (74.0-90.0)	85.0 (76.0-95.0)	66.0 (60.0-73.0)	69.0 (62.0-77.0)
Uric acid (umol/l)	339.2 (64.9)	353.2 (73.4)	260.4 (62.9)	293.5 (80.1)
Glucose (mmol/l)	4.1 (3.6-4.8)	4.3 (3.7-5.1)	4.2 (3.7-4.9)	4.3 (3.7-5.3)
HbA1c (%)	5.4 (5.3-5.6)	5.5 (5.3-5.8)	5.4 (5.3-5.6)	5.6 (5.3-5.8)
ALP(iU/l)	67.5 (58.0-78.0)	72.5 (61.0-87.0)	64.0 (51.0-79.0)	74.0 (62.0-88.0)
ALT (iU/l)	23.0 (18.0-30.0)	23.0 (18.0-29.0)	18.0 (14.0-23.5)	19.0 (15.0-24.0)
AST (iU/l)	31.0 (27.0-35.0)	31.0 (28.0-36.0)	28.0 (25.0-32.0)	29.0 (26.0-34.0)
GGT (iU/l)	25.0 (19.0-39.0)	28.0 (21.0-42.0)	17.0 (13.0-26.0)	20.5 (15.0-31.0)
Calcium (mmol/l)	2.5 (0.1)	2.5 (0.1)	2.5 (0.1)	2.5 (0.1)
Magnesium (mmol/l)	0.9 (0.1)	0.9 (0.1)	0.9 (0.1)	0.9 (0.1)
Iron (umol/l)	18.6 (5.3)	18.0 (5.5)	17.1 (5.9)	16.9 (5.4)
Transferrin (umol/l)	34.5 (5.0)	34.4 (4.7)	36.5 (5.7)	35.9 (5.2)
Ferritin (pmol/l)	249.4 (143.8-422.5)	274.2 (159.6-457.8)	119.1 (62.9-207.8)	168.5 (105.6-267.4)
Total bilirubin (umol/l)	7.0 (5.0-10.0)	7.0 (5.0-10.0)	6.0 (5.0-8.0)	6.0 (4.0-7.0)

Netherlands

	Males		Females	
	No CHD	**CHD**	**No CHD**	**CHD**
N	147	347	959	840
Age (years)	45.8 (8.7)	51.6 (6.9)	55.3 (8.0)	58.3 (7.4)
Current smoker	60 (40.8%)	160 (46.1%)	225 (23.5%)	274 (32.6%)
BMI (kg/m2)	25.9 (3.1)	26.7 (3.5)	25.3 (4.0)	26.3 (4.2)
Diabetes	1 (0.7%)	7 (2.0%)	19 (2.0%)	47 (5.6%)

Table S3: Continued

Netherlands

	Males		Females	
	No CHD	**CHD**	**No CHD**	**CHD**
Hypertension	39 (26.5%)	182 (52.4%)	368 (38.4%)	497 (59.2%)
SBP (mmHg)	124.8 (14.8)	133.1 (18.8)	129.0 (19.5)	137.1 (21.6)
Use of antihypertensive medication at baseline	7 (4.8%)	35 (10.1%)	178 (18.6%)	272 (32.4%)
Use of lipid lowering medication at baseline	0 (0.0%)	10 (2.9%)	40 (4.2%)	75 (8.9%)
Total cholesterol (mmol/l)	6.0 (1.2)	6.5 (1.2)	6.2 (1.1)	6.7 (1.2)
HDL cholesterol (mmol/l)	1.2 (0.3)	1.2 (0.3)	1.6 (0.4)	1.4 (0.4)
non-HDL cholesterol (mmol/l)	4.7 (1.3)	5.3 (1.2)	4.7 (1.2)	5.2 (1.2)
Triglycerides (mmol/l)	1.5 (0.9-2.0)	1.6 (1.1-2.4)	1.2 (0.9-1.6)	1.4 (1.0-2.1)
ApoA1 (g/l)	1.4 (0.2)	1.4 (0.2)	1.6 (0.3)	1.6 (0.3)
ApoB (g/l)	1.1 (0.3)	1.2 (0.3)	1.0 (0.3)	1.2 (0.3)
Lp(a) (mg/dl)	48.0 (26.0-69.5)	49.5 (27.0-82.3)	37.4 (20.6-68.4)	48.0 (24.9-87.5)
CRP (mg/l)	1.0 (0.5-2.2)	1.6 (0.7-3.3)	1.3 (0.6-3.0)	1.9 (0.9-3.9)
Albumin (g/l)	48.3 (3.3)	47.7 (3.1)	46.2 (2.8)	46.1 (3.3)
Creatinine (umol/l)	79.0 (74.0-86.0)	80.0 (72.0-89.0)	65.0 (59.0-72.0)	66.0 (59.0-72.0)
Uric acid (umol/l)	349.4 (65.4)	357.8 (74.7)	264.0 (63.8)	279.5 (66.3)
Glucose (mmol/l)	4.1 (3.5-4.7)	4.3 (3.8-5.1)	4.1 (3.7-4.7)	4.3 (3.8-5.1)
HbA1c (%)	5.3 (5.1-5.5)	5.4 (5.3-5.7)	5.4 (5.3-5.6)	5.6 (5.3-5.8)
ALP(iU/l)	64.0 (57.0-78.0)	73.0 (62.0-82.5)	68.0 (56.0-82.0)	73.0 (60.0-86.0)
ALT (iU/l)	25.0 (17.0-36.0)	26.0 (20.0-34.0)	17.0 (14.0-23.0)	18.0 (15.0-24.0)
AST (iU/l)	29.0 (25.0-36.0)	29.0 (26.0-34.0)	26.0 (23.0-30.0)	26.0 (23.0-30.0)
GGT (iU/l)	28.0 (19.0-47.0)	31.0 (23.0-46.5)	18.0 (14.0-25.0)	20.0 (15.0-28.0)
Calcium (mmol/l)	2.5 (0.1)	2.5 (0.1)	2.5 (0.1)	2.5 (0.1)
Magnesium (mmol/l)	0.9 (0.1)	0.9 (0.1)	0.9 (0.1)	0.9 (0.1)
Iron (umol/l)	20.3 (5.3)	20.1 (6.3)	17.7 (5.8)	17.1 (5.2)

Table S3: Continued

Netherlands				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
Transferrin (umol/l)	33.5 (4.3)	34.2 (4.5)	34.5 (5.0)	34.1 (4.9)
Ferritin (pmol/l)	465.2 (261.8-674.2)	456.2 (267.4-691.0)	204.5 (110.1-340.5)	238.2 (141.0-379.8)
Total bilirubin (umol/l)	8.0 (6.0-11.0)	8.0 (6.0-10.0)	6.0 (5.0-8.0)	6.0 (4.0-7.0)

Greece				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
N	374	188	622	84
Age (years)	51.3 (11.8)	56.5 (10.8)	52.5 (11.4)	63.7 (7.3)
Current smoker	161 (43.0%)	97 (51.6%)	130 (20.9%)	10 (11.9%)
BMI (kg/m2)	27.9 (3.9)	28.5 (3.7)	28.4 (5.3)	29.7 (5.2)
Diabetes	19 (5.1%)	32 (17.0%)	31 (5.0%)	23 (27.4%)
Hypertension	128 (34.2%)	97 (51.6%)	237 (38.1%)	58 (69.0%)
SBP (mmHg)	129.8 (17.9)	136.3 (19.6)	128.5 (21.6)	146.1 (20.0)
Use of antihypertensive medication at baseline	56 (15.0%)	45 (23.9%)	128 (20.6%)	40 (47.6%)
Use of lipid lowering medication at baseline	16 (4.3%)	19 (10.1%)	35 (5.6%)	16 (19.0%)
Total cholesterol (mmol/l)	6.1 (1.2)	6.5 (1.3)	6.0 (1.1)	6.7 (1.3)
HDL cholesterol (mmol/l)	1.2 (0.3)	1.1 (0.3)	1.5 (0.4)	1.4 (0.4)
non-HDL cholesterol (mmol/l)	4.9 (1.2)	5.4 (1.3)	4.5 (1.1)	5.3 (1.4)
Triglycerides (mmol/l)	1.3 (0.9-1.8)	1.6 (1.1-2.3)	1.0 (0.7-1.5)	1.4 (1.0-1.8)
ApoA1 (g/l)	1.4 (0.2)	1.3 (0.2)	1.6 (0.3)	1.6 (0.3)
ApoB (g/l)	1.1 (0.3)	1.3 (0.3)	1.0 (0.3)	1.2 (0.3)
Lp(a) (mg/dl)	40.9 (22.5-69.4)	59.1 (39.0-82.3)	38.2 (20.4-57.9)	62.0 (47.3-91.8)
CRP (mg/l)	1.2 (0.6-2.6)	1.7 (0.8-3.6)	1.3 (0.6-2.8)	1.9 (1.1-3.7)
Albumin (g/l)	46.7 (2.8)	46.8 (2.6)	45.9 (2.8)	45.9 (3.2)

Table S3: Continued

Greece				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
Creatinine (umol/l)	73.0 (66.0-82.0)	72.0 (64.0-81.0)	56.0 (50.0-63.0)	54.0 (48.8-62.0)
Uric acid (umol/l)	336.5 (75.8)	335.1 (73.9)	250.6 (67.4)	268.6 (80.7)
Glucose (mmol/l)	4.0 (3.4-4.7)	4.1 (3.5-5.1)	3.9 (3.3-4.5)	4.3 (3.6-5.9)
HbA1c (%)	5.5 (5.3-5.9)	5.8 (5.4-6.3)	5.6 (5.3-5.9)	6.0 (5.7-7.0)
ALP(iU/l)	61.0 (51.0-72.0)	64.0 (54.0-77.0)	63.0 (51.0-76.0)	69.5 (58.0-82.2)
ALT (iU/l)	22.0 (17.0-29.0)	23.0 (18.8-30.0)	17.0 (14.0-22.0)	18.0 (15.0-22.0)
AST (iU/l)	28.0 (24.0-33.0)	29.0 (26.0-33.2)	26.0 (22.0-30.0)	26.0 (22.0-29.0)
GGT (iU/l)	21.0 (16.0-30.8)	25.0 (18.0-35.2)	13.0 (11.0-17.0)	15.0 (12.0-20.0)
Calcium (mmol/l)	2.5 (0.1)	2.5 (0.1)	2.5 (0.1)	2.5 (0.1)
Magnesium (mmol/l)	0.9 (0.1)	0.9 (0.1)	0.9 (0.1)	0.8 (0.1)
Iron (umol/l)	18.6 (6.2)	17.4 (5.9)	15.9 (6.4)	15.9 (5.7)
Transferrin (umol/l)	33.9 (4.9)	33.4 (4.8)	35.4 (5.4)	34.3 (5.1)
Ferritin (pmol/l)	253.9 (148.9-393.3)	241.6 (146.1-417.9)	98.9 (47.2-166.3)	150.6 (98.9-220.2)
Total bilirubin (umol/l)	8.5 (6.0-12.0)	7.5 (6.0-10.0)	7.0 (5.0-9.0)	7.0 (5.0-9.0)

Germany				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
N	496	214	785	83
Age (years)	51.8 (8.0)	55.8 (7.0)	48.8 (8.9)	55.9 (7.7)
Current smoker	121 (24.4%)	95 (44.4%)	134 (17.1%)	28 (33.7%)
BMI (kg/m2)	26.6 (3.3)	27.6 (3.5)	25.3 (4.4)	27.4 (5.2)
Diabetes	23 (4.6%)	30 (14.0%)	22 (2.8%)	9 (10.8%)
Hypertension	274 (55.2%)	156 (72.9%)	306 (39.0%)	53 (63.9%)
SBP (mmHg)	134.5 (17.4)	140.5 (18.0)	124.8 (17.7)	136.5 (20.3)
Use of antihypertensive medication at baseline	144 (29.0%)	80 (37.4%)	176 (22.4%)	44 (53.0%)

Table S3: Continued

Germany				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
Use of lipid lowering medication at baseline	60 (12.1%)	38 (17.8%)	57 (7.3%)	12 (14.5%)
Total cholesterol (mmol/l)	5.9 (1.1)	6.3 (1.1)	5.8 (1.1)	6.5 (1.1)
HDL cholesterol (mmol/l)	1.3 (0.4)	1.2 (0.3)	1.7 (0.4)	1.4 (0.4)
non-HDL cholesterol (mmol/l)	4.6 (1.1)	5.1 (1.1)	4.1 (1.1)	5.0 (1.2)
Triglycerides (mmol/l)	1.4 (1.0-2.0)	1.8 (1.2-2.5)	1.0 (0.7-1.4)	1.3 (0.9-2.0)
ApoA1 (g/l)	1.5 (0.2)	1.4 (0.2)	1.7 (0.3)	1.6 (0.3)
ApoB (g/l)	1.0 (0.3)	1.1 (0.3)	0.9 (0.2)	1.1 (0.3)
Lp(a) (mg/dl)	36.9 (20.6-64.6)	51.2 (24.3-89.2)	28.6 (14.3-57.4)	37.3 (17.0-97.8)
CRP (mg/l)	1.0 (0.5-2.0)	1.7 (0.8-3.5)	1.1 (0.5-2.2)	1.8 (0.9-3.2)
Albumin (g/l)	46.9 (2.7)	46.4 (2.7)	46.2 (2.9)	46.0 (2.4)
Creatinine (umol/l)	79.0 (71.0-88.2)	80.0 (72.0-90.0)	63.0 (56.0-70.0)	63.0 (56.0-72.5)
Uric acid (umol/l)	344.3 (73.2)	355.3 (75.3)	244.5 (59.2)	275.2 (81.6)
Glucose (mmol/l)	5.3 (4.8-5.8)	5.2 (4.9-6.1)	5.0 (4.6-5.5)	5.4 (4.8-6.0)
HbA1c (%)	5.4 (5.2-5.6)	5.6 (5.3-6.0)	5.3 (5.1-5.5)	5.5 (5.3-6.0)
ALP(iU/l)	64.0 (55.0-76.0)	66.0 (56.0-79.0)	56.0 (46.0-69.0)	68.0 (52.0-84.5)
ALT (iU/l)	25.0 (19.0-35.0)	23.0 (18.0-32.0)	15.0 (12.0-21.0)	18.0 (13.0-23.0)
AST (iU/l)	29.0 (25.0-34.0)	28.0 (24.0-33.0)	23.0 (21.0-27.0)	26.0 (21.0-30.0)
GGT (iU/l)	31.0 (21.0-50.0)	35.0 (23.0-50.0)	17.0 (13.0-25.0)	21.0 (16.0-33.0)
Calcium (mmol/l)	2.4 (0.1)	2.4 (0.1)	2.4 (0.1)	2.4 (0.1)
Magnesium (mmol/l)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)
Iron (umol/l)	17.5 (6.1)	17.6 (5.8)	16.3 (6.1)	16.2 (4.5)
Transferrin (umol/l)	32.8 (4.6)	32.2 (4.3)	34.8 (5.3)	32.9 (4.9)
Ferritin (pmol/l)	471.9 (283.1-705.6)	474.1 (301.7-826.4)	152.8 (76.4-271.9)	262.9 (126.9-402.2)
Total bilirubin (umol/l)	8.0 (6.0-10.0)	7.0 (5.0-9.0)	6.0 (5.0-8.0)	6.0 (4.5-7.0)

Table S3: Continued

Sweden				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
N	436	662	814	423
Age (years)	58.1 (7.1)	61.4 (6.8)	57.0 (8.2)	62.6 (7.1)
Current smoker	121 (27.8%)	258 (39.0%)	220 (27.0%)	163 (38.5%)
BMI (kg/m2)	25.6 (3.5)	26.4 (3.7)	25.0 (4.1)	26.6 (4.7)
Diabetes	0 (0.0%)	49 (7.4%)	0 (0.0%)	40 (9.5%)
Hypertension	218 (50.0%)	483 (73.0%)	369 (45.3%)	312 (73.8%)
SBP (mmHg)	141.8 (18.1)	151.9 (20.0)	139.2 (20.4)	153.2 (22.0)
Use of antihypertensive medication at baseline	71 (16.3%)	209 (31.6%)	137 (16.8%)	142 (33.6%)
Use of lipid lowering medication at baseline	7 (1.6%)	30 (4.5%)	12 (1.5%)	21 (5.0%)
Total cholesterol (mmol/l)	6.1 (1.0)	6.3 (1.1)	6.3 (1.2)	6.9 (1.2)
HDL cholesterol (mmol/l)	1.3 (0.4)	1.2 (0.4)	1.6 (0.4)	1.5 (0.4)
non-HDL cholesterol (mmol/l)	4.8 (1.0)	5.1 (1.1)	4.6 (1.2)	5.4 (1.2)
Triglycerides (mmol/l)	1.4 (1.1-2.1)	1.7 (1.2-2.4)	1.2 (0.9-1.7)	1.5 (1.1-2.2)
ApoA1 (g/l)	1.4 (0.2)	1.3 (0.2)	1.6 (0.3)	1.5 (0.3)
ApoB (g/l)	1.1 (0.2)	1.2 (0.3)	1.0 (0.3)	1.2 (0.3)
Lp(a) (mg/dl)	36.5 (20.9-62.8)	45.0 (25.3-77.4)	29.1 (16.1-57.1)	43.1 (24.0-82.4)
CRP (mg/l)	1.2 (0.6-2.4)	1.9 (0.9-4.1)	1.1 (0.6-2.5)	2.1 (1.0-4.8)
Albumin (g/l)	46.1 (3.0)	45.6 (2.8)	45.3 (2.7)	45.4 (2.7)
Creatinine (umol/l)	80.0 (73.0-88.0)	79.5 (72.0-88.0)	64.0 (58.0-70.0)	64.0 (56.0-71.0)
Uric acid (umol/l)	333.9 (67.8)	351.4 (74.7)	257.9 (60.4)	285.6 (75.9)
Glucose (mmol/l)	5.5 (5.0-6.1)	5.7 (5.2-6.5)	5.3 (4.9-5.7)	5.5 (5.1-6.1)
HbA1c (%)	5.7 (5.4-5.9)	5.8 (5.5-6.2)	5.6 (5.4-5.9)	5.9 (5.6-6.3)
ALP(iU/l)	68.5 (58.8-80.0)	74.0 (62.0-86.0)	64.0 (54.0-79.0)	75.0 (62.0-91.0)
ALT (iU/l)	22.0 (17.0-29.0)	21.0 (16.0-30.0)	16.0 (12.0-21.0)	17.0 (13.0-23.5)
AST (iU/l)	30.0 (26.0-34.0)	29.0 (26.0-35.0)	26.0 (23.0-30.0)	27.0 (24.0-31.0)

Table S3: Continued

Sweden

	Males		Females	
	No CHD	**CHD**	**No CHD**	**CHD**
GGT (iU/l)	29.0 (21.0-44.0)	32.5 (23.0-51.8)	19.0 (15.0-30.0)	24.0 (18.0-38.0)
Calcium (mmol/l)	2.4 (0.1)	2.4 (0.1)	2.4 (0.1)	2.4 (0.1)
Magnesium (mmol/l)	0.9 (0.1)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)
Iron (umol/l)	17.9 (5.6)	17.4 (5.8)	16.8 (5.6)	16.2 (5.4)
Transferrin (umol/l)	32.8 (4.4)	33.0 (4.7)	33.9 (4.8)	33.9 (4.6)
Ferritin (pmol/l)	385.4 (234.9-564.0)	377.6 (213.5-620.2)	168.5 (94.9-289.3)	222.5 (132.6-333.7)
Total bilirubin (umol/l)	7.0 (5.0-10.0)	7.0 (5.0-9.0)	6.0 (5.0-8.0)	6.0 (4.0-7.0)

Denmark

	Males		Females	
	No CHD	**CHD**	**No CHD**	**CHD**
N	790	915	756	366
Age (years)	56.5 (4.3)	57.9 (4.3)	56.5 (4.4)	59.1 (4.1)
Current smoker	264 (33.4%)	451 (49.3%)	262 (34.7%)	191 (52.2%)
BMI (kg/m2)	26.4 (3.4)	27.2 (3.7)	25.3 (4.3)	26.5 (4.5)
Diabetes	16 (2.0%)	45 (4.9%)	5 (0.7%)	14 (3.8%)
Hypertension	397 (50.3%)	603 (65.9%)	358 (47.4%)	259 (70.8%)
SBP (mmHg)	140.3 (19.2)	147.4 (21.4)	136.7 (20.1)	148.8 (23.2)
Use of antihypertensive medication at baseline	81 (10.3%)	167 (18.3%)	90 (11.9%)	103 (28.1%)
Use of lipid lowering medication at baseline	7 (0.9%)	30 (3.3%)	2 (0.3%)	14 (3.8%)
Total cholesterol (mmol/l)	5.9 (1.0)	6.1 (1.1)	5.9 (1.1)	6.3 (1.2)
HDL cholesterol (mmol/l)	1.4 (0.4)	1.2 (0.4)	1.7 (0.5)	1.5 (0.4)
non-HDL cholesterol (mmol/l)	4.5 (1.1)	4.9 (1.1)	4.3 (1.1)	4.9 (1.3)
Triglycerides (mmol/l)	1.4 (1.0-2.0)	1.7 (1.2-2.3)	1.1 (0.8-1.5)	1.4 (1.0-1.9)
ApoA1 (g/l)	1.4 (0.3)	1.3 (0.2)	1.6 (0.3)	1.5 (0.3)
ApoB (g/l)	1.0 (0.2)	1.1 (0.2)	0.9 (0.2)	1.1 (0.3)

Table S3: Continued

Denmark				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
Lp(a) (mg/dl)	41.3 (21.9-73.1)	46.3 (26.6-82.6)	31.6 (17.2-62.7)	48.1 (19.2-97.8)
CRP (mg/l)	1.1 (0.5-2.4)	1.7 (0.8-3.6)	1.1 (0.6-2.5)	2.0 (0.9-4.7)
Albumin (g/l)	44.6 (3.5)	44.1 (3.4)	43.9 (3.6)	43.5 (3.5)
Creatinine (umol/l)	79.0 (71.0-86.0)	78.0 (70.5-88.0)	63.0 (57.0-70.0)	65.0 (58.0-72.0)
Uric acid (umol/l)	334.2 (72.8)	346.2 (79.1)	247.8 (63.8)	276.2 (70.7)
Glucose (mmol/l)	5.3 (5.0-5.8)	5.4 (5.0-6.0)	5.2 (4.8-5.6)	5.3 (4.9-5.8)
HbA1c (%)	5.4 (5.2-5.7)	5.6 (5.3-5.9)	5.4 (5.3-5.6)	5.6 (5.3-5.9)
ALP(iU/l)	64.0 (53.0-77.0)	69.0 (58.0-81.0)	61.0 (49.0-74.0)	70.0 (56.0-84.0)
ALT (iU/l)	21.0 (16.0-29.0)	21.0 (16.0-30.5)	15.0 (12.0-20.0)	17.0 (13.0-22.0)
AST (iU/l)	29.0 (25.0-34.0)	28.0 (24.5-35.0)	26.0 (22.8-29.0)	26.0 (22.0-31.0)
GGT (iU/l)	30.5 (22.0-53.0)	33.0 (23.0-56.5)	19.0 (14.0-31.0)	23.0 (16.0-36.0)
Calcium (mmol/l)	2.3 (0.1)	2.3 (0.1)	2.3 (0.1)	2.3 (0.1)
Magnesium (mmol/l)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)
Iron (umol/l)	17.3 (5.9)	16.6 (5.8)	15.8 (5.3)	15.6 (5.8)
Transferrin (umol/l)	31.7 (4.6)	31.8 (4.6)	32.5 (5.1)	32.3 (4.4)
Ferritin (pmol/l)	370.8 (218.0-576.4)	359.6 (192.2-565.1)	170.8 (92.1-283.1)	197.8 (119.6-291.6)
Total bilirubin (umol/l)	7.0 (5.0-9.0)	6.0 (5.0-9.0)	6.0 (4.0-7.0)	5.0 (4.0-7.0)

Oxford				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
N	31	77	119	134
Age (years)	51.5 (12.8)	59.7 (9.1)	48.8 (10.5)	60.1 (8.9)
Current smoker	1 (3.2%)	7 (9.1%)	7 (5.9%)	8 (6.0%)
BMI (kg/m2)	23.4 (2.5)	26.1 (3.5)	23.8 (3.9)	25.6 (4.9)
Diabetes	0 (0.0%)	6 (7.8%)	0 (0.0%)	13 (9.7%)

Table S3: Continued

Oxford				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
Hypertension	9 (29.0%)	40 (51.9%)	19 (16.0%)	75 (56.0%)
SBP (mmHg)	128.3 (16.0)	140.0 (18.5)	121.6 (16.6)	139.2 (22.8)
Use of antihypertensive medication at baseline	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Use of lipid lowering medication at baseline	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Total cholesterol (mmol/l)	5.5 (1.0)	5.9 (1.0)	5.7 (1.3)	6.3 (1.2)
HDL cholesterol (mmol/l)	1.4 (0.4)	1.3 (0.3)	1.7 (0.4)	1.5 (0.4)
non-HDL cholesterol (mmol/l)	4.1 (0.9)	4.7 (1.0)	4.0 (1.3)	4.8 (1.3)
Triglycerides (mmol/l)	1.1 (0.9-1.7)	1.5 (1.1-2.0)	0.9 (0.7-1.4)	1.4 (1.0-2.2)
ApoA1 (g/l)	1.5 (0.3)	1.4 (0.2)	1.6 (0.3)	1.6 (0.3)
ApoB (g/l)	0.9 (0.2)	1.0 (0.2)	0.9 (0.3)	1.1 (0.3)
Lp(a) (mg/dl)	39.4 (21.8-81.3)	45.4 (30.5-71.5)	42.1 (23.4-72.2)	50.0 (30.6-83.1)
CRP (mg/l)	0.5 (0.2-0.7)	1.1 (0.6-2.7)	0.7 (0.3-1.6)	1.6 (0.6-4.5)
Albumin (g/l)	48.4 (2.3)	46.9 (2.5)	47.2 (2.5)	46.1 (3.1)
Creatinine (umol/l)	80.0 (75.0-85.5)	81.0 (73.0-92.0)	62.0 (57.0-66.0)	63.0 (57.0-71.0)
Uric acid (umol/l)	327.4 (68.7)	329.5 (65.8)	243.1 (52.9)	272.2 (66.2)
Glucose (mmol/l)	1.9 (0.6-3.2)	2.6 (1.6-3.5)	2.3 (1.4-3.4)	2.5 (1.4-3.6)
HbA1c (%)	5.2 (5.0-5.3)	5.3 (5.2-5.6)	5.3 (5.1-5.4)	5.4 (5.2-5.7)
ALP(iU/l)	66.0 (53.5-76.5)	70.0 (60.0-84.0)	63.0 (54.0-75.0)	74.0 (62.0-86.0)
ALT (iU/l)	24.0 (18.5-27.0)	21.0 (17.0-28.0)	17.0 (14.0-21.0)	19.0 (15.0-24.0)
AST (iU/l)	33.0 (30.5-37.5)	34.0 (31.0-39.0)	31.0 (27.0-34.5)	32.0 (28.0-37.0)
GGT (iU/l)	21.0 (17.0-27.5)	26.0 (18.0-37.0)	14.0 (12.0-19.0)	19.0 (15.0-26.0)
Calcium (mmol/l)	2.5 (0.1)	2.5 (0.1)	2.5 (0.1)	2.5 (0.2)
Magnesium (mmol/l)	0.9 (0.1)	0.9 (0.1)	0.9 (0.1)	0.9 (0.1)
Iron (umol/l)	20.7 (6.4)	21.0 (5.4)	17.9 (5.5)	19.1 (6.9)
Transferrin (umol/l)	35.6 (3.4)	34.4 (4.0)	38.1 (5.3)	36.8 (5.5)

Table S3: Continued

Oxford				
	Males		**Females**	
	No CHD	**CHD**	**No CHD**	**CHD**
Ferritin (pmol/l)	206.7 (118.0-306.8)	202.2 (132.6-411.2)	71.9 (42.7-130.3)	165.2 (87.6-260.7)
Total bilirubin (umol/l)	11.0 (7.0-13.0)	9.0 (7.0-12.0)	7.0 (5.0-9.0)	7.0 (5.0-9.0)

Values represent N (%), mean (standard deviation), or median (25^{th} - 75^{th} percentile). CHD: coronary heart disease, SBP: systolic blood pressure, HDL: high-density lipoprotein, apo: apolipoprotein, Lp(a): lipoprotein(a), CRP: C-reactive protein, HbA1c: glycated hemoglobin, ALP: alkaline phosphatase, ALT: alanine transaminase, AST: aspartate aminotransferase, GGT: gamma-glutamyl transferase.

Table S4: Comparison of baseline characteristics of participants with complete data for all predictors and with at least one missing value for the predictors.

	Males				Females			
	No CHD		CHD		No CHD		CHD	
	Complete cases	Missings	Complete cases	Missings	Complete cases	Missings	Complete cases	Missings
N	3307	2337	3931	2722	6524	3370	2937	1547
Age (years)	53.6 (8.6)	51.0 (8.4)	58.5 (7.9)	56.1 (7.9)	53.5 (8.9)	49.2 (9.0)	60.0 (7.6)	58.6 (8.4)
Current smoker	1023 (30.9%)	757 (33.0%)	1530 (38.9%)	1119 (41.5%)	1435 (22.0%)	724 (21.9%)	900 (30.6%)	418 (27.3%)
BMI (kg/m2)	26.4 (3.5)	27.3 (3.7)	27.0 (3.6)	27.7 (3.8)	25.6 (4.5)	26.7 (4.9)	26.7 (4.6)	27.4 (4.8)
Diabetes	103 (3.1%)	120 (5.1%)	281 (7.1%)	221 (8.1%)	124 (1.9%)	105 (3.1%)	218 (7.4%)	150 (9.7%)
Hypertension	1458 (44.1%)	656 (28.1%)	2501 (63.6%)	1306 (48.0%)	2450 (37.6%)	844 (25.0%)	1894 (64.5%)	877 (56.7%)
SBP (mmHg)	135.3 (18.3)	133.7 (18.3)	143.7 (20.3)	143.0 (19.9)	129.9 (19.8)	128.6 (19.5)	142.6 (22.3)	144.6 (21.3)
Use of antihypertensive medication at baseline	434 (13.1%)	281 (12.0%)	932 (23.7%)	380 (14.0%)	670 (10.3%)	483 (14.3%)	770 (26.2%)	377 (24.4%)
Use of lipid lowering medication at baseline	105 (3.2%)	53 (2.3%)	200 (5.1%)	71 (2.6%)	159 (2.4%)	89 (2.6%)	178 (6.1%)	75 (4.8%)
Total cholesterol (mmol/l)	5.9 (1.1)	5.9 (1.1)	6.3 (1.1)	6.3 (1.1)	6.0 (1.1)	5.8 (1.1)	6.6 (1.2)	6.5 (1.4)
HDL cholesterol (mmol/l)	1.3 (0.4)	1.3 (0.3)	1.2 (0.3)	1.2 (0.3)	1.6 (0.4)	1.6 (0.4)	1.5 (0.4)	1.4 (0.4)

Table S4: Continued

	Males				Females			
	No CHD		CHD		No CHD		CHD	
	Complete cases	Missings	Complete cases	Missings	Complete cases	Missings	Complete cases	Missings
non-HDL cholesterol (mmol/l)	4.6 (1.1)	4.6 (1.1)	5.1 (1.1)	5.2 (1.2)	4.4 (1.2)	4.2 (1.2)	5.2 (1.3)	5.1 (1.4)
Triglycerides (mmol/l)	1.3 (1.0-1.9)	1.2 (0.9-1.9)	1.7 (1.2-2.3)	1.6 (1.1-2.5)	1.0 (0.8-1.5)	0.9 (0.7-1.3)	1.4 (1.0-2.0)	1.4 (1.0-2.3)
ApoA1 (g/l)	1.4 (0.2)	1.4 (0.3)	1.4 (0.2)	1.4 (0.3)	1.6 (0.3)	1.6 (0.3)	1.6 (0.3)	1.6 (0.3)
ApoB (g/l)	1.0 (0.2)	1.1 (0.3)	1.2 (0.3)	1.2 (0.2)	1.0 (0.3)	0.9 (0.3)	1.2 (0.3)	1.1 (0.3)
Lp(a) (mg/dl)	41.4 (22.5-70.3)	45.3 (26.5-72.8)	50.2 (28.4-83.3)	56.4 (32.8-92.6)	34.9 (18.3-62.4)	38.7 (20.8-70.0)	49.4 (25.9-94.0)	51.8 (28.3-96.5)
CRP (mg/l)	1.0 (0.6-2.1)	1.1 (0.6-2.1)	1.7 (0.8-3.6)	1.6 (0.8-3.1)	1.1 (0.5-2.5)	1.1 (0.5-2.3)	1.9 (0.9-4.1)	1.9 (0.9-4.0)
Albumin (g/l)	46.4 (3.2)	47.0 (3.0)	45.9 (3.2)	46.6 (3.2)	45.8 (3.1)	46.0 (2.9)	45.6 (3.3)	45.7 (3.2)
Creatinine (umol/l)	78.0 (71.0-86.0)	78.0 (71.0-86.0)	80.0 (72.0-90.0)	78.0 (70.0-87.0)	63.0 (56.0-70.0)	61.0 (55.0-68.0)	65.0 (58.0-72.0)	62.0 (55.0-70.0)
Uric acid (umol/l)	336.4 (71.2)	343.7 (75.8)	349.2 (75.4)	352.3 (78.8)	252.0 (62.8)	249.3 (62.3)	279.9 (72.9)	280.7 (72.1)
Glucose (mmol/l)	5.1 (4.5-5.7)	5.1 (4.5-5.7)	5.1 (4.3-5.9)	5.2 (4.5-6.0)	4.8 (4.1-5.3)	4.8 (4.3-5.3)	4.9 (4.0-5.6)	5.0 (4.4-5.8)
HbA1c (%)	5.4 (5.3-5.7)	5.4 (5.2-5.6)	5.6 (5.3-5.9)	5.6 (5.3-5.9)	5.4 (5.3-5.7)	5.4 (5.2-5.6)	5.6 (5.3-6.0)	5.6 (5.3-6.0)
ALP(iU/l)	64.0 (55.0-76.0)	62.0 (52.0-73.0)	70.0 (59.0-83.0)	67.0 (57.0-81.0)	62.0 (50.0-76.0)	60.0 (48.0-73.0)	73.0 (60.0-87.0)	71.0 (58.0-89.0)
ALT (iU/l)	22.0 (17.0-31.0)	24.0 (19.0-33.0)	23.0 (17.0-30.0)	25.0 (19.0-33.0)	17.0 (13.0-22.0)	17.0 (13.0-22.0)	18.0 (14.0-24.0)	20.0 (16.0-26.0)

Table S4: Continued

	Males				Females			
	No CHD		CHD		No CHD		CHD	
	Complete cases	Missings	Complete cases	Missings	Complete cases	Missings	Complete cases	Missings
AST (iU/l)	29.0 (25.0-34.0)	29.0 (25.0-35.0)	30.0 (26.0-35.0)	30.0 (25.0-35.0)	26.0 (22.0-30.0)	25.0 (22.0-29.0)	27.0 (24.0-32.0)	28.0 (24.0-33.0)
GGT (iU/l)	28.0 (20.0-44.0)	27.0 (19.2-42.0)	31.0 (22.0-47.0)	32.0 (23.0-49.0)	17.0 (13.0-24.0)	15.0 (12.0-21.0)	20.0 (15.0-31.0)	20.0 (15.0-31.0)
Calcium (mmol/l)	2.4 (0.1)	2.5 (0.1)	2.4 (0.1)	2.5 (0.1)	2.4 (0.1)	2.5 (0.1)	2.5 (0.1)	2.5 (0.2)
Magnesium (mmol/l)	0.8 (0.1)	0.9 (0.1)	0.9 (0.1)	0.9 (0.1)	0.8 (0.1)	0.8 (0.1)	0.8 (0.1)	0.9 (0.1)
Iron (umol/l)	18.1 (5.8)	18.1 (6.2)	17.8 (5.8)	17.9 (6.3)	16.7 (6.0)	16.0 (6.6)	16.8 (5.6)	15.9 (5.8)
Transferrin (umol/l)	33.3 (4.8)	34.6 (4.9)	33.4 (4.8)	34.6 (5.4)	34.9 (5.5)	36.1 (5.5)	34.5 (5.2)	35.0 (5.2)
Ferritin (pmol/l)	337.1 (188.8-556.2)	276.4 (152.8-485.4)	332.6 (186.5-548.3)	316.9 (159.6-548.3)	148.3 (74.2-260.7)	89.9 (44.9-168.5)	206.7 (119.1-328.1)	159.6 (89.9-276.4)
Total bilirubin (umol/l)	8.0 (6.0-11.0)	8.0 (6.0-11.0)	7.0 (5.0-10.0)	7.0 (5.0-10.0)	6.0 (5.0-9.0)	7.0 (5.0-9.0)	6.0 (4.0-8.0)	6.0 (4.0-8.0)
France	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	510 (7.8%)	57 (1.7%)	36 (1.2%)	4 (0.3%)
Italy	573 (17.3%)	68 (2.9%)	432 (11.0%)	60 (2.2%)	1150 (17.6%)	147 (4.4%)	354 (12.1%)	45 (2.9%)
Spain	154 (4.7%)	1173 (50.2%)	90 (2.3%)	794 (29.2%)	354 (5.4%)	1821 (54.0%)	55 (1.9%)	288 (18.6%)
UK	306 (9.3%)	33 (1.4%)	1006 (25.6%)	205 (7.5%)	455 (7.0%)	42 (1.2%)	562 (19.1%)	125 (8.1%)
Netherlands	147 (4.4%)	20 (0.9%)	347 (8.8%)	81 (3.0%)	959 (14.7%)	61 (1.8%)	840 (28.6%)	240 (15.5%)
Greece	374 (11.3%)	55 (2.4%)	188 (4.8%)	42 (1.5%)	622 (9.5%)	94 (2.8%)	84 (2.9%)	18 (1.2%)

Table S4: Continued

	Males				Females			
	No CHD		CHD		No CHD		CHD	
	Complete cases	Missings	Complete cases	Missings	Complete cases	Missings	Complete cases	Missings
Germany	496 (15.0%)	319 (13.6%)	214 (5.4%)	273 (10.0%)	785 (12.0%)	380 (11.3%)	83 (2.8%)	58 (3.7%)
Sweden	436 (13.2%)	592 (25.3%)	662 (16.8%)	808 (29.7%)	814 (12.5%)	669 (19.9%)	423 (14.4%)	331 (21.4%)
Denmark	790 (23.9%)	64 (2.7%)	915 (23.3%)	240 (8.8%)	756 (11.6%)	66 (2.0%)	366 (12.5%)	124 (8.0%)
Oxford	31 (0.9%)	13 (0.6%)	77 (2.0%)	219 (8.0%)	119 (1.8%)	33 (1.0%)	134 (4.6%)	314 (20.3%)

Values represent N (%), mean (standard deviation), or median (25^{th} - 75^{th} percentile). CHD: coronary heart disease, SBP: systolic blood pressure, HDL: high-density lipoprotein, apo: apolipoprotein, Lp(a): lipoprotein(a), CRP: C-reactive protein, HbA1c: glycated hemoglobin, ALP: alkaline phosphatase, ALT: alanine transaminase, AST: aspartate aminotransferase, GGT: gamma-glutamyl transferase.

Table S5: Other performance measures of the traditional prediction model, traditional prediction models with one biomarker added, the combined model and the full model.

	Males				Females			
	R^2	**BIC**	**AIC**	**Brier score**	**R^2**	**BIC**	**AIC**	**Brier score**
TP	0.296	79291.3	79290.7	0.0502	0.266	54747.8	54747.1	0.0203
TP + Non-HDL cholesterol	0.298	79249.0	79247.2	0.0502	0.267	54744.5	54742.7	0.0203
TP + Log triglycerides	0.298	79254.4	79252.7	0.0501	0.266	54749.6	54747.8	0.0203
TP + ApoA1	0.296	79281.8	79280.0	0.0502	0.266	54750.3	54748.5	0.0203
TP + ApoB	0.307	79098.8	79097.0	0.0500	0.270	54679.3	54677.6	0.0203
TP + Sqrt Lp(A)	0.299	79236.0	79234.2	0.0502	0.269	54702.0	54700.3	0.0203
TP + Log CRP	0.304	79140.8	79139.0	0.0501	0.275	54586.5	54584.7	0.0202
TP + Albumin	0.299	79236.0	79234.2	0.0502	0.269	54703.9	54702.1	0.0203
TP + Log creatinine	0.296	79290.8	79289.1	0.0502	0.266	54746.7	54745.0	0.0203
TP + Uric acid	0.296	79293.9	79292.1	0.0502	0.267	54734.0	54732.2	0.0203
TP + Glucose	0.297	79264.8	79263.1	0.0502	0.268	54725.9	54724.1	0.0203
TP + HbA1c	0.305	79120.5	79118.7	0.0501	0.272	54640.6	54638.8	0.0202
TP + ALP	0.297	79270.5	79268.7	0.0504	0.267	54742.9	54741.2	0.0203
TP + ALT	0.296	79290.8	79289.0	0.0502	0.266	54749.4	54747.7	0.0203
TP + AST	0.296	79283.8	79282.0	0.0502	0.266	54750.3	54748.5	0.0203
TP + GGT	0.296	79294.0	79292.3	0.0502	0.267	54738.0	54736.2	0.0203
TP + Calcium	0.297	79267.7	79265.9	0.0502	0.266	54750.5	54748.7	0.0203
TP + Magnesium	0.296	79288.3	79286.6	0.0502	0.267	54743.1	54741.3	0.0203
TP + Iron	0.297	79268.0	79266.3	0.0502	0.266	54746.9	54745.1	0.0203
TP + Transferrin	0.297	79273.5	79271.7	0.0502	0.266	54750.9	54749.1	0.0203
TP + Log ferritin	0.296	79279.1	79277.3	0.0502	0.266	54750.2	54748.4	0.0203
TP + Total bilirubin	0.296	79293.9	79292.1	0.0502	0.266	54749.7	54748.0	0.0203
Full	0.334	78691.3	78651.6	0.0500	0.288	54404.2	54364.6	0.0202
Combined	0.327	78765.8	78747.8	0.0501	0.285	54399.7	54387.4	0.0202

TP (traditional prediction) model includes age, current smoking, diabetes, hypertension, log bmi, systolic blood pressure, log total cholesterol, log HDL cholesterol. Full model includes all predictors from the traditional prediction model plus predictors listed in the table. Combined model includes all predictors from the traditional prediction model plus log triglycerides, apoB, log CRP,

albumin, HbA1c, ALP, ALT, and iron in males, and apoB, sqrt Lp(a), log CRP, albumin, HbA1c, AST, and magnesium in females. BIC: Bayesian Information Criterion, AIC: Akaike Information Criterion, HDL: high-density lipoprotein, apo: apolipoprotein, Lp(a): lipoprotein(a), CRP: C-reactive protein, HbA1c: glycated hemoglobin, ALP: alkaline phosphatase, ALT: alanine transaminase, AST: aspartate aminotransferase, GGT: gamma-glutamyl transferase.

Table S6: Performance of the traditional prediction model and the traditional prediction models with one biomarker added stratified by country.

Males, c-statistic										
	France	**Italy**	**Spain**	**UK**	**Nether-lands**	**Greece**	**Ger-many**	**Sweden**	**Den-mark**	**Oxford**
TP	NA	0.669	0.663	0.609	0.620	0.692	0.707	0.630	0.644	0.618
TP + Non-HDL cholesterol	NA	0.669	0.662	0.610	0.619	0.693	0.707	0.631	0.645	0.619
TP + Log triglycerides	NA	0.668	0.664	0.610	0.619	0.691	0.707	0.631	0.646	0.620
TP + ApoA1	NA	0.669	0.664	0.610	0.618	0.693	0.707	0.631	0.644	0.619
TP + ApoB	NA	0.677	0.665	0.607	0.617	0.693	0.705	0.631	0.648	0.619
TP + Sqrt Lp(A)	NA	0.670	0.669	0.611	0.623	0.696	0.708	0.631	0.643	0.619
TP + Log CRP	NA	0.673	0.666	0.611	0.622	0.697	0.706	0.632	0.646	0.619
TP + Albumin	NA	0.671	0.664	0.608	0.620	0.690	0.708	0.633	0.648	0.620
TP + Log creatinine	NA	0.669	0.663	0.609	0.620	0.692	0.707	0.630	0.645	0.618
TP + Uric acid	NA	0.669	0.663	0.609	0.620	0.692	0.707	0.630	0.644	0.618
TP + Glucose	NA	0.668	0.662	0.610	0.622	0.691	0.706	0.630	0.644	0.620
TP + HbA1c	NA	0.668	0.666	0.611	0.622	0.692	0.708	0.630	0.650	0.622
TP + ALP	NA	0.669	0.665	0.610	0.620	0.691	0.707	0.630	0.645	0.618
TP + ALT	NA	0.668	0.663	0.609	0.621	0.692	0.707	0.630	0.644	0.619
TP + AST	NA	0.668	0.663	0.609	0.621	0.692	0.707	0.630	0.645	0.619
TP + GGT	NA	0.669	0.663	0.609	0.620	0.692	0.707	0.630	0.644	0.618
TP + Calcium	NA	0.672	0.663	0.609	0.619	0.692	0.709	0.631	0.645	0.619
TP + Magnesium	NA	0.668	0.663	0.609	0.621	0.691	0.706	0.630	0.645	0.619
TP + Iron	NA	0.668	0.663	0.609	0.620	0.695	0.707	0.630	0.646	0.620

Table S6: Continued

Males, c-statistic										
	France	**Italy**	**Spain**	**UK**	**Nether-lands**	**Greece**	**Ger-many**	**Sweden**	**Den-mark**	**Oxford**
TP + Transferrin	NA	0.667	0.664	0.608	0.620	0.691	0.705	0.630	0.644	0.618
TP + Log ferritin	NA	0.666	0.663	0.608	0.621	0.691	0.707	0.630	0.646	0.618
TP + Total bilirubin	NA	0.669	0.663	0.609	0.620	0.692	0.706	0.630	0.645	0.619
Full	NA	0.680	0.678	0.613	0.621	0.698	0.710	0.638	0.657	0.623

Females, c-statistic										
	France	**Italy**	**Spain**	**UK**	**Nether-lands**	**Greece**	**Ger-many**	**Sweden**	**Den-mark**	**Oxford**
TP	0.782	0.720	0.761	0.658	0.632	0.806	0.800	0.733	0.706	0.653
TP + Non-HDL cholesterol	0.781	0.720	0.761	0.658	0.632	0.806	0.800	0.733	0.706	0.653
TP + Log triglycerides	0.782	0.720	0.761	0.658	0.632	0.806	0.801	0.733	0.706	0.653
TP + ApoA1	0.782	0.720	0.761	0.658	0.632	0.807	0.801	0.733	0.705	0.653
TP + ApoB	0.780	0.724	0.762	0.659	0.632	0.809	0.799	0.734	0.710	0.653
TP + Sqrt Lp(A)	0.783	0.720	0.761	0.661	0.632	0.814	0.802	0.734	0.708	0.653
TP + Log CRP	0.785	0.721	0.760	0.665	0.637	0.806	0.804	0.735	0.714	0.655
TP + Albumin	0.782	0.719	0.761	0.657	0.632	0.805	0.802	0.732	0.710	0.655
TP + Log creatinine	0.784	0.719	0.761	0.658	0.632	0.806	0.800	0.732	0.707	0.653
TP + Uric acid	0.787	0.719	0.761	0.659	0.632	0.806	0.802	0.733	0.708	0.653
TP + Glucose	0.782	0.719	0.761	0.658	0.633	0.809	0.801	0.732	0.706	0.654
TP + HbA1c	0.791	0.723	0.763	0.659	0.634	0.812	0.803	0.734	0.710	0.654
TP + ALP	0.781	0.720	0.761	0.659	0.632	0.806	0.800	0.732	0.708	0.652
TP + ALT	0.782	0.720	0.761	0.658	0.632	0.806	0.800	0.733	0.706	0.653
TP + AST	0.782	0.720	0.761	0.658	0.632	0.806	0.800	0.733	0.706	0.653
TP + GGT	0.783	0.720	0.761	0.659	0.632	0.807	0.801	0.733	0.706	0.653

Table S6: Continued

Females, c-statistic

	France	Italy	Spain	UK	Nether-lands	Greece	Ger-many	Sweden	Den-mark	Oxford
TP + Calcium	0.782	0.719	0.761	0.658	0.632	0.806	0.800	0.733	0.706	0.653
TP + Magnesium	0.783	0.719	0.760	0.658	0.631	0.807	0.801	0.732	0.706	0.653
TP + Iron	0.783	0.719	0.761	0.658	0.632	0.807	0.801	0.733	0.706	0.653
TP + Transferrin	0.781	0.719	0.761	0.658	0.632	0.806	0.801	0.733	0.706	0.653
TP + Log ferritin	0.781	0.720	0.761	0.659	0.632	0.806	0.801	0.732	0.706	0.653
TP + Total bilirubin	0.782	0.719	0.761	0.658	0.632	0.806	0.800	0.733	0.706	0.653
Full	0.806	0.726	0.764	0.666	0.638	0.817	0.808	0.740	0.724	0.657

Males, OE ratio

	France	Italy	Spain	UK	Nether-lands	Greece	Ger-many	Sweden	Den-mark	Oxford
TP	NA	1.341	1.192	0.801	1.249	0.911	1.034	0.870	0.858	1.402
TP + Non-HDL cholesterol	NA	1.345	1.202	0.805	1.248	0.907	1.026	0.866	0.869	1.394
TP + Log triglycerides	NA	1.304	1.173	0.808	1.267	0.874	1.053	0.869	0.876	1.444
TP + ApoA1	NA	1.346	1.205	0.802	1.260	0.918	1.056	0.856	0.849	1.422
TP + ApoB	NA	1.294	1.139	0.830	1.287	0.877	1.069	0.843	0.902	1.433
TP + Sqrt Lp(A)	NA	1.345	1.183	0.796	1.262	0.926	1.037	0.872	0.858	1.382
TP + Log CRP	NA	1.335	1.197	0.800	1.228	0.922	1.033	0.871	0.861	1.447
TP + Albumin	NA	1.354	1.225	0.808	1.270	0.918	1.046	0.854	0.817	1.478
TP + Log creatinine	NA	1.342	1.192	0.800	1.247	0.918	1.034	0.871	0.859	1.399
TP + Uric acid	NA	1.342	1.190	0.801	1.248	0.912	1.034	0.871	0.858	1.402
TP + Glucose	NA	1.328	1.199	0.826	1.272	0.936	1.014	0.851	0.846	1.513
TP + HbA1c	NA	1.337	1.259	0.817	1.260	0.875	1.022	0.830	0.868	1.430

Table S6: Continued

Males, OE ratio										
	France	**Italy**	**Spain**	**UK**	**Nether-lands**	**Greece**	**Ger-many**	**Sweden**	**Den-mark**	**Oxford**
TP + ALP	NA	1.348	1.199	0.790	1.246	0.915	1.036	0.870	0.855	1.394
TP + ALT	NA	1.337	1.189	0.803	1.250	0.905	1.039	0.873	0.857	1.408
TP + AST	NA	1.328	1.189	0.808	1.249	0.903	1.037	0.871	0.858	1.431
TP + GGT	NA	1.342	1.193	0.801	1.249	0.915	1.033	0.870	0.856	1.403
TP + Calcium	NA	1.356	1.227	0.807	1.259	0.923	1.037	0.858	0.825	1.449
TP + Magnesium	NA	1.339	1.195	0.805	1.257	0.914	1.033	0.869	0.850	1.441
TP + Iron	NA	1.352	1.192	0.805	1.280	0.917	1.025	0.867	0.848	1.452
TP + Transferrin	NA	1.327	1.175	0.791	1.259	0.911	1.050	0.874	0.872	1.369
TP + Log ferritin	NA	1.334	1.175	0.791	1.276	0.890	1.064	0.877	0.865	1.390
TP + Total bilirubin	NA	1.348	1.193	0.801	1.250	0.913	1.033	0.870	0.856	1.405
Full	NA	1.297	1.222	0.811	1.311	0.841	1.082	0.806	0.891	1.462

Females, OE ratio										
	France	**Italy**	**Spain**	**UK**	**Nether-lands**	**Greece**	**Ger-many**	**Sweden**	**Den-mark**	**Oxford**
TP	1.074	0.942	1.245	0.574	0.641	0.584	1.001	0.580	0.605	1.322
TP + Non-HDL cholesterol	1.075	0.942	1.245	0.574	0.642	0.584	1.002	0.580	0.605	1.321
TP + Log triglycerides	1.069	0.936	1.238	0.575	0.643	0.581	1.004	0.581	0.608	1.330
TP + ApoA1	1.077	0.948	1.242	0.574	0.641	0.585	1.009	0.574	0.601	1.320
TP + ApoB	1.061	0.940	1.234	0.584	0.649	0.584	1.017	0.567	0.623	1.334
TP + Sqrt Lp(A)	1.081	0.945	1.223	0.568	0.641	0.592	0.999	0.586	0.603	1.294
TP + Log CRP	1.075	0.929	1.264	0.571	0.622	0.599	1.002	0.572	0.580	1.318
TP + Albumin	1.064	0.961	1.271	0.573	0.648	0.584	1.010	0.563	0.563	1.364
TP + Log creatinine	1.072	0.945	1.251	0.570	0.637	0.597	1.003	0.578	0.605	1.318

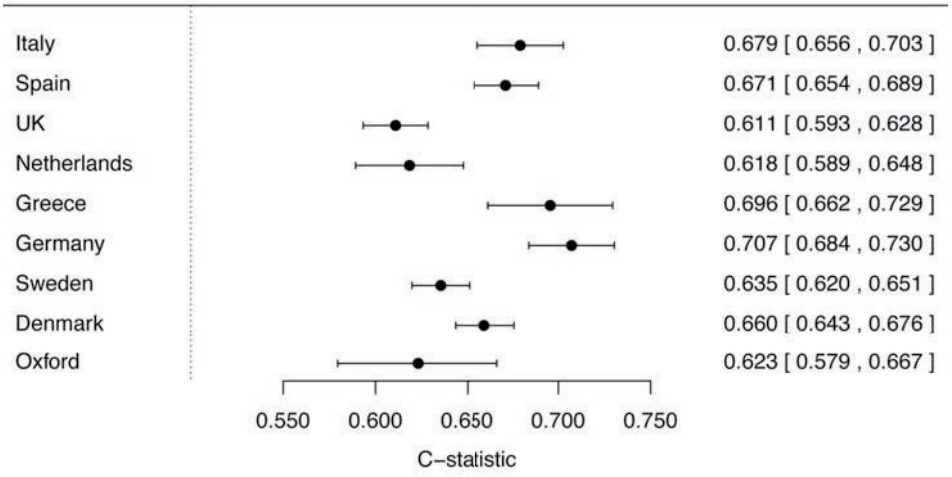
Table S6: Continued

Females, OE ratio

	France	Italy	Spain	UK	Nether-lands	Greece	Ger-many	Sweden	Den-mark	Oxford
TP + Uric acid	1.062	0.946	1.252	0.570	0.636	0.591	1.006	0.578	0.607	1.322
TP + Glucose	1.098	0.926	1.245	0.579	0.650	0.603	0.988	0.571	0.594	1.418
TP + HbA1c	1.065	0.922	1.258	0.581	0.646	0.572	1.021	0.562	0.616	1.365
TP + ALP	1.086	0.935	1.244	0.572	0.637	0.588	1.007	0.580	0.603	1.305
TP + ALT	1.074	0.942	1.245	0.574	0.641	0.583	1.001	0.580	0.605	1.322
TP + AST	1.074	0.942	1.245	0.574	0.641	0.583	1.001	0.580	0.605	1.322
TP + GGT	1.053	0.946	1.268	0.571	0.640	0.604	0.992	0.573	0.593	1.317
TP + Calcium	1.076	0.940	1.240	0.573	0.640	0.583	1.001	0.581	0.609	1.322
TP + Magnesium	1.080	0.937	1.248	0.575	0.643	0.584	0.994	0.578	0.595	1.369
TP + Iron	1.082	0.945	1.240	0.575	0.643	0.586	1.000	0.579	0.601	1.332
TP + Transferrin	1.072	0.944	1.248	0.576	0.640	0.584	1.001	0.578	0.602	1.329
TP + Log ferritin	1.066	0.940	1.253	0.576	0.637	0.589	0.995	0.579	0.604	1.328
TP + Total bilirubin	1.073	0.939	1.245	0.574	0.642	0.582	1.001	0.580	0.607	1.321
Full	1.014	0.930	1.266	0.575	0.619	0.585	1.071	0.536	0.591	1.328

TP (traditional prediction) model includes age, current smoking, diabetes, hypertension, log bmi, systolic blood pressure, log total cholesterol, log HDL cholesterol. Full model includes all predictors from the traditional prediction model plus predictors listed in the table. Marked in grey: top 3 biomarkers with incremental value per country. HDL: high-density lipoprotein, apo: apolipoprotein, Lp(a): lipoprotein(a), CRP: C-reactive protein, HbA1c: glycated hemoglobin, ALP: alkaline phosphatase, ALT: alanine transaminase, AST: aspartate aminotransferase, GGT: gamma-glutamyl transferase, NA: not applicable.

Males



Females

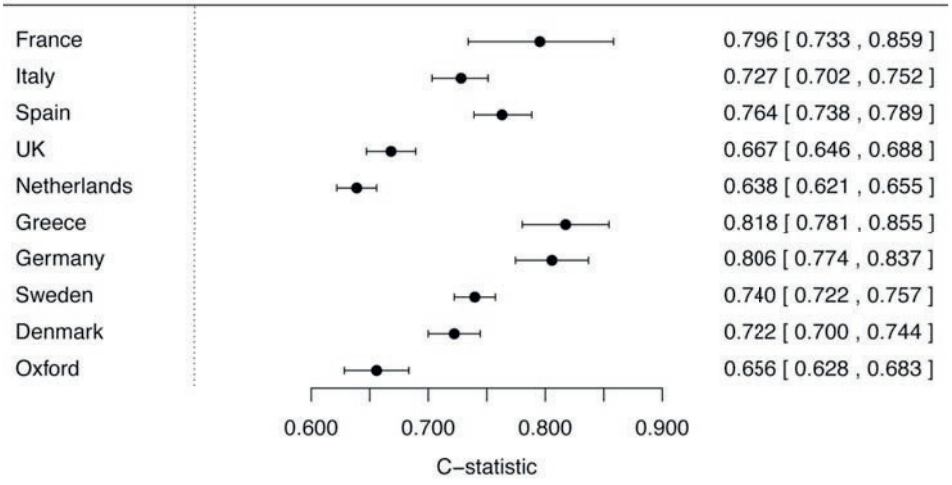
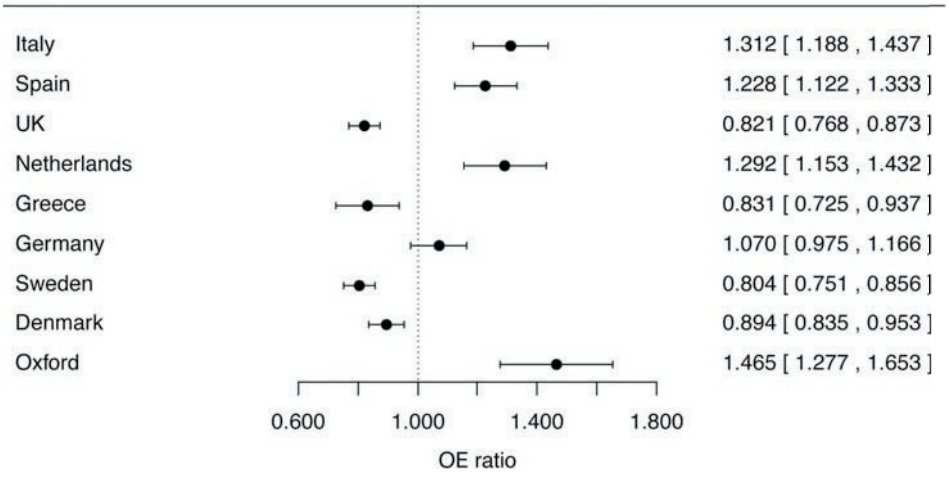


Figure S1: C-statistic of combined model, stratified by country.

Males



Females

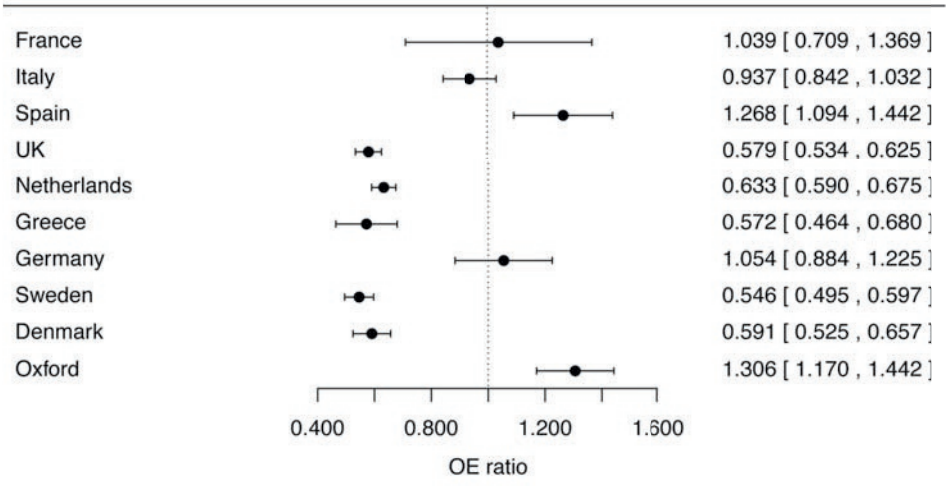


Figure S2: OE ratio of combined model, stratified by country.

Chapter 6

Treatment use in prognostic model research: a systematic review of cardiovascular prognostic studies

Romin Pajouheshnia
Johanna AAG Damen
Rolf HH Groenwold
Karel GM Moons
Linda M Peelen

Diagn Progn Res. 2017;1(1):15

Abstract

Background: Ignoring treatments in prognostic model development or validation can affect the accuracy and transportability of models. We aim to quantify the extent to which the effects of treatment have been addressed in existing prognostic model research and provide recommendations for the handling and reporting of treatment use in future studies.

Methods: We first describe how and when the use of treatments by individuals in a prognostic study can influence the development or validation of a prognostic model. We subsequently conducted a systematic review of the handling and reporting of treatment use in prognostic model studies in cardiovascular medicine. Data on treatment use (e.g. medications, surgeries, lifestyle interventions), timing of their use, and the handling of such treatment use in the analyses were extracted and summarized.

Results: 302 articles were included in the review. Treatment use was not mentioned in 91 (30%) articles. 146 (48%) reported specific information about treatment use in their studies; 78 (26%) provided information about multiple treatments. Three articles (1%) reported changes in medication use ("treatment drop-in") during follow-up. 79 articles (26%) excluded treated individuals from their analysis, 80 articles (26%) modelled treatment as an outcome and of the 155 articles that developed a model, 86 (55%) modelled treatment use, almost exclusively at baseline, as a predictor.

Conclusions: The use of treatments has been partly considered by the majority of CVD prognostic model studies. Detailed accounts including, for example, information on treatment drop-in were rare. Where relevant, the use of treatments should be considered in the analysis of prognostic model studies, particularly when a prognostic model is designed to guide the use of certain treatments and these treatments have been used by the study participants. Future prognostic model studies should clearly report the use of treatments by study participants and consider the potential impact of treatment use on the study findings.

Background

An important part of prognostic research is the development and validation of prognostic models or risk scores. These models can be used to make individualised predictions of a person's absolute risk of developing a specific health outcome[1,2] and can, for example, be used to inform different aspects of clinical decision making. A notable example of this is in cardiovascular medicine: if a patient's risk of a cardiovascular event is predicted to be above a specific probability threshold, lifestyle changes are recommended, with or without initiation of preventative medication.[3-5]

Concerns have been raised that the use of treatments, such as pharmacological therapy or diet and lifestyle-related interventions, may have an unwanted impact when patient data (e.g. from a cohort or registry) is used to develop or validate a prognostic model.[6-8] In order to develop or validate prognostic models that predict an individual's probability of developing an outcome in the absence of a certain treatment (i.e. their untreated health course) one should ideally include people who have not received that treatment before or during follow-up.[1,6] In practice, however, such prognostic models are often derived from or validated in data sets where a proportion of the individuals has received that specific treatment. If, for example, treatments were administered in a study according to individuals' predicted risks (either implicitly or explicitly), a model developed using this data will likely underestimate the risk of the predicted outcome in the absence of treatment, and could thus lead to under-treatment when such a model is used in future individuals.[8,9]

In this manuscript we aim to provide insight into the problems that arise when treatment use is ignored when developing or validating a prognostic model. First, we elaborate on how and when treatment use could negatively impact prognostic modelling. Following this, we provide evidence of the scale of this issue in published studies by means of a systematic literature review of the reporting and handing of treatment use in cardiovascular prognostic model research. We conclude with suggestions for the handling and reporting of treatment use in prognostic model research.

Methods

What do we mean by "treatment" and when is it a problem?

Herein, we use "treatment" to refer to any intervention, medical (e.g. medication, surgery, therapy) or non-medical (e.g. quit smoking or do more exercise), undertaken by an individual that lowers their risk of a certain outcome. We also include in this definition modifications that an individual makes to their behaviour or lifestyle that reduce their risk of a specific outcome. We propose two categories of treatment: "guided" and "background". The term "guided treatments" refers to treatments that one intends to

guide or direct by means of the prognostic model being developed or validated. For example, CVD prediction models are used to guide the prescription of lipid-lowering medication, as well as direct targeted advice about lifestyle changes to high-risk individuals. "Background treatments" refer to any other treatment that an individual receives during a prognostic study. This could, for example, include treatments that are part of routine medical care or changes an individual makes to their lifestyle. Figure 1 outlines the different stages where treatments may be used in a prognostic study.

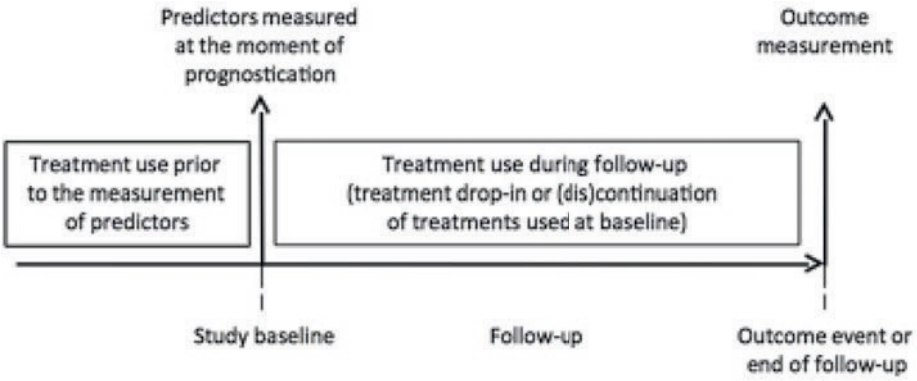


Figure 1: The timing of treatment use in a prognostic study

Guided treatments

Prognostic models are often used to guide or direct the initiation of certain treatments or interventions. In this case a prognostic model should estimate the risk of developing a certain outcome *if individuals were to remain untreated* with this particular treatment (so-called "untreated risk prediction").[1,8,10] If this particular, "guided" treatment is given to study participants after the predictors are measured but before the ascertainment of the outcome (henceforth we refer to this as "treatment drop-in", see Figure 1), the chance of treated individuals developing the outcome of interest will be decreased. Crucially, the outcomes measured in the study will no longer represent the untreated outcomes that the model is designed to predict. It follows that models developed using data from individuals who received guided treatments will provide biased underestimates of (untreated) risks in future individuals, if treatment use is ignored.[8] In validation studies, models will incorrectly appear to overestimate risk if applied in individuals that receive the specific guided treatment.[8,11]

Background treatments

Participants in a prognostic study commonly receive risk-lowering treatments during follow-up as a part of routine care. As in the case of guided treatments, if these

"background" treatments are effective in lowering the risk of the outcome under prediction, we can expect a reduction in the probability of treated individuals developing the outcome of interest. However, unlike with guided treatments, the outcomes measured in the study still reflect the outcome under prediction. Background treatments should instead be considered to be a part of the case-mix of participants in study. Provided the pattern of treatment use, and the effect of the treatment on the outcome risk is consistent across populations, differences between model performance in the development cohort new populations should not be due to treatment use. However, background treatment use and effectiveness, may vary between settings. For example, a model developed in a setting where everyone received some standard (effective) treatment during follow-up may not be transportable to a different population where that intervention is not available, or a less effective alternative treatment is routinely used. In this case the predicted probabilities provided by the model in this new population will be too low.

Examples

We illustrate the distinction between different types of treatment with two hypothetical examples, from two different clinical domains.

Example 1: A model is developed to predict six-month mortality risk in patients with end-stage renal disease (ERD) in the absence of a kidney transplantation. The model will be used to help decide which future patients will receive a kidney transplant. In the development cohort, all patients began risk-lowering haemodialysis after enrolment as a part of routine care and a subset of patients additionally received a kidney transplant.

Example 2: A validation study is conducted to evaluate an existing prognostic model for the prediction of five-year CVD risk in the general population. The model is used in practice to decide whether lipid-lowering drugs (statins) will be prescribed. Several individuals in the study were prescribed risk-lowering statins and were recommended to modify their lifestyle based on their predicted CVD risk. In addition, a number of patients took other risk-lowering medications (e.g. aspirin) as a part of routine care.

In both examples, some study participants initiated one or more treatments or interventions after predictor measurements were taken. In example 1, we can consider haemodialysis to be a "background" treatment, as described above, which requires no further consideration for model development. However, the model may need to be recalibrated for settings where haemodialysis is not a part of usual care or where a substantial proportion of patients receive some other type of (e.g. peritoneal) dialysis. In contrast, kidney transplant, a treatment guided by predictions made by the model, could bias model development. The outcomes measured in individuals who received a transplant during follow-up do not reflect our outcome of interest: six-month mortality

without kidney transplantation. Not taking this into account in model development will lead to a prediction model that actually underestimates the risk of mortality without transplantation in future patients with ERD.

In example 2, the use of medications such as aspirin can be considered as background treatment that will not affect the validity of the validation study. It may however explain model miscalibration in the validation cohort if the pattern of use or the effectiveness of these treatments is different from those of the development cohort. With regard to lipid-lowering medication, ideally one would validate the model in individuals who have not received lipid-lowering medication during follow-up. As high-risk individuals received statins in the study, their risk of a CVD event in the study is lower than it would have been, had they remained untreated. In this example lifestyle changes merit separate attention. If the model is used in practice, as with statins, to help target lifestyle advice to high-risk individuals, this treatment should not be ignored in the validation study. However, many individuals may have modified their lifestyles independent of any targeted advice, in which case, lifestyle changes could be viewed as a background treatment.

To summarize, when treatments are initiated in participants after the moment of prognostication (see Figure 1), the risk-lowering effects of these treatments may impact on model development or validation. We propose that the intended use and thus kind of risk predictions a model aims to provide (i.e. prognosis with or without treatment), as well as the types of treatments (guided or background) used in a data set or study, are key factors that determine how treatments may impact on prognostic model development or validation. For further details on the challenges of treatment use and how to account for them in prognostic model development and validation, see Groenwold et al.[8] and, Pajouheshnia et al.[11] respectively, and further guidance can be found in Figure 4.

A review of treatment use in published prognostic model studies

To provide insight into the extent to which treatment use has been addressed in the development and validation of prognostic models, we used a previously conducted systematic review of the reporting and analysis of prognostic models for predicting the risk of the future occurrence of CVD outcomes in the general population.[12] A completed PRISMA checklist for this review is found in Additional File 1.

Data sources, search and study selection

In brief, a search was performed on 1st June 2013 in MEDLINE and Embase to identify original research articles reporting the development (derivation of a new model) or external validation (evaluation of an existing model in a new population) of a prognostic model, and "incremental value studies", in which the additional value of a certain predictor or (bio)marker was assessed on top of either an existing risk score or a model consisting of a core set of conventional predictors (e.g. age, sex, smoking, systolic blood pressure, cholesterol, diabetes).

Titles and abstracts were first screened for eligibility, and subsequent full-text screening was conducted. Publications were considered for inclusion if they were original articles that reported cardiovascular risk prognostic modelling in a general population setting. Full details of the search strategy and in-/exclusion criteria can be found in the original review.[12]

Data extraction

Directed by the CHARMS checklist,[13] a list of key items (Additional file 2) for extraction was derived for the current review by one author (RP) and updated after group consideration (RP, LMP, RHHG, JAAGD, KGMM). As the aim of this review is to provide an overview of research practice and reporting, study quality and risk of bias assessment was not conducted. Independent data extraction was piloted among three authors (RP, JAAGD, RHHG). The remaining data extraction was conducted by one author (RP) and any queries were discussed primarily with one author (JAAGD), and then two other authors (LMP, RHHG) until a consensus was reached.

General study characteristics were extracted for each article, including the study design used to collect data, the start and end dates of participant data collection and the prediction horizons of reported models. Relevant treatments or interventions for cardiovascular disease prevention were defined prior to data extraction and broadly divided into three classes: pharmacological treatments (notably antihypertensive, lipid-lowering and antithrombotic medication), cardiovascular surgical interventions (e.g. coronary revascularization, carotid endarterectomy), and lifestyle interventions. While the term "lifestyle interventions" can refer to changes in a diverse range of modifiable risk factors, we defined this in our review as the reporting of active modifications to exercise, nutritional or smoking habits, as a part of a programme or following physician recommendations. All reported information on treatment use and how it was considered in the analysis was extracted (for full details, see Additional file 2).

Results of the literature review

General characteristics of included articles

The search of the original systematic review identified 9965 unique records, of which 1388 were found to be relevant following title and abstract screening, as previously reported.[12] After full text screening for eligibility, 302 articles were included for review (Additional file 3). A summary of the article inclusion process is presented in Figure 2.

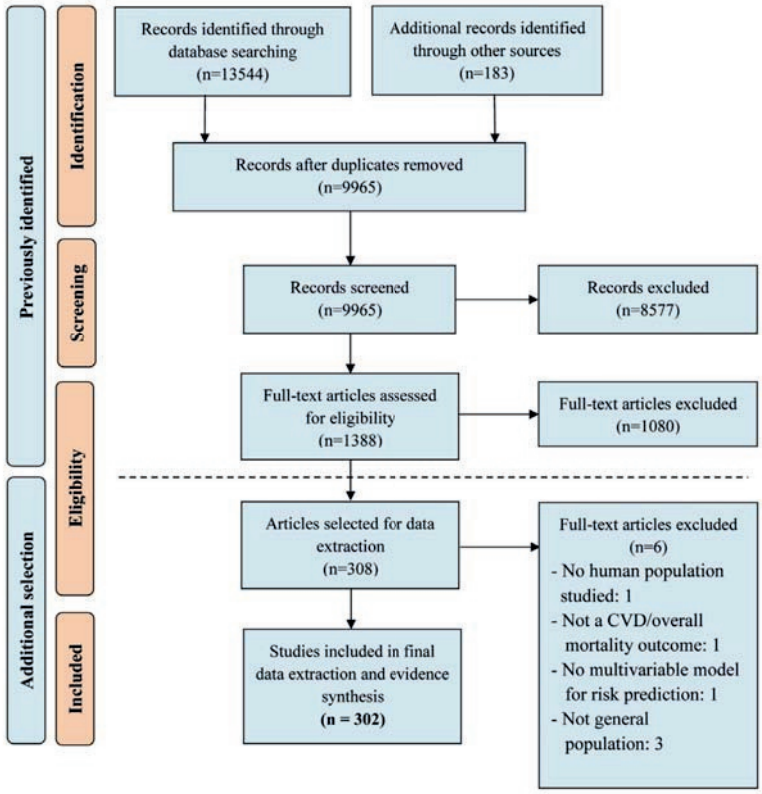


Figure 2: A flow diagram of article inclusion and exclusion.

The final set of articles includes publications from 102 different journals. Publication dates ranged from 1967 to 2013 and 157 articles (52%) were published from 2009 onwards. Participant data collection ranged from as early as 1948 until 2011. Further details are presented in Table 1.

Table 1: General characteristics of the included articles

Characteristics of included studies (n = 302)	
Study type	
Development	124
Validation	146
Incremental value assessment	135
Over a set of core predictors	81

Table 1: Continued

Characteristics of included studies (n = 302)	
Design of study used for prognostic modelling	
Observational	286
Randomized trial	16
Follow-up period (years)	10, (6, 12); 15% †
Prediction horizon (years)	10, (8, 10); 12% †

* One article may have multiple study types (e.g. the development and validation of a model); thus values do not sum to the total number of included articles. † Values represent as follows: median, (lower quartile, upper quartile); percentage of studies that did not report this information.

Reporting and handling of treatment use

Overall, nearly one-third (91 articles, 30%) of the 302 included articles did not report any information about relevant preventative or therapeutic treatments. The reporting of treatments in prognostic modelling articles has increased over time, as illustrated in Figure 3. Just over half of the articles published up until 2008 (81 articles, 56%) reported information about treatment, whereas from 2009 to June 2013 this increased (130 articles, 83%). Summaries of the reporting and handling of information about treatment use are presented in Table 2 and Table 3, respectively.

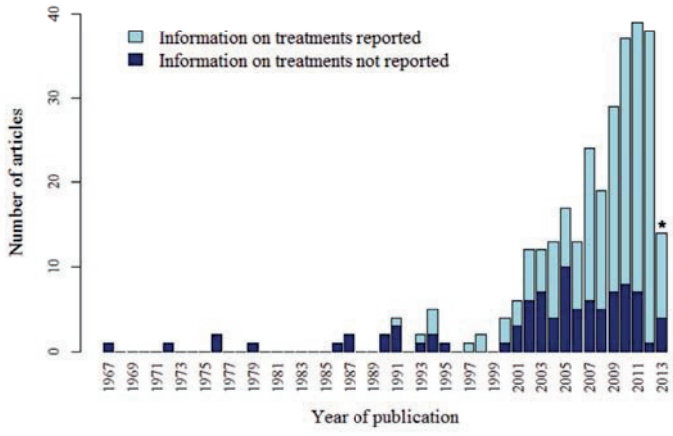


Figure 3: Reporting of treatment in CVD prognostic modelling studies over time. Articles were classified as having reported information on treatment if the use of at least one potentially risk-lowering treatment in the study was reported, or if the effect of a treatment on the study findings was discussed. (*) Articles were included up to June 2013; this column only represents treatment reporting during the first half of 2013.

Development studies

Of the 124 articles that reported the development of a new prognostic model, baseline information on treatment use was reported in 43 articles (35%). Six articles (5%) reported treatment use during follow-up, two (2%) reported changes in medication use during follow-up, four (3%) described incident surgical procedures (cardiovascular surgeries occurring after the study baseline) and in 11 articles (9%), the timing of treatments was unclear. Two articles reported that information on treatment was not available. Treatment use was most often accounted for in analyses by modelling treatment as a predictor (54 articles, 44%). 20 articles (15%) excluded treated individuals from the analysis. Changes in treatment use during follow-up were not modelled.

Incremental value studies

In articles that reported the evaluation of the incremental value of a predictor over either a core set of predictors or an existing model, baseline information about treatment use was reported for 74 articles (55%). Changes in medication use were reported in three articles, and surgical procedures that occurred during follow-up were reported in 15 articles (11%). Five articles (4%) reported that information on treatment use was not available. Where incremental value was assessed over a set of core predictors, treatment use was accounted for most often by including treatment as one of the core predictors (48 articles, 59%). 53 articles (39%) excluded treated individuals from analyses. Surgical outcomes were frequently modelled as a part of a composite endpoint (58 articles, 43%).

Validation studies

In studies that externally validated (evaluated) an existing CVD prognostic model, where reported, most information about treatment use was measured at baseline only (55 articles, 37%). No articles reported changes in medication use during follow-up. Four articles reported a lack of available data on treatment use. In addition, five articles (3%) presented information about treatment use in the population in which the model was originally developed, of which two reported differences of more than 10% in the proportion of baseline treatment users between the development study and the validation study. Another five articles (3%) commented on how differences between treatment use in the development and validation populations could have contributed to poor performance of the model upon validation. Medication use was accounted for exclusively by restricting analyses to untreated patients (38 articles, 26%). In addition, 35 articles (24%) accounted for incident surgical procedures by including surgery within the composite endpoint of their study.

Table 2: Reporting of treatment use by study type

Reported treatment	Overall (n = 302) (%) *	Development studies (n = 124) (%)	Incremental value studies (n = 135) (%)	Validation studies (n = 146) (%)
Medication use (any)	135 (45)	45 (36)	73 (54)	62 (41)
Antihypertensive	122 (41)	40 (32)	66 (49)	58 (38)
Lipid-lowering	81 (27)	24 (19)	47 (33)	38 (26)
Antithrombotic/ anticoagulant	17 (6)	2 (2)	15 (11)	7 (5)
Lifestyle interventions	2 (1)	1 (1)	0	1 (1)
Surgical interventions	39 (13)	9 (7)	26 (19)	15 (11)

* One article may have multiple study types (e.g. the development and validation of a model); thus values in individual columns do not sum to the overall number of included articles. Articles may have reported multiple treatments and thus percentages in each column should not necessarily sum to 100%.

Table 3: Handling of treatment in the analyses of prognostic model studies

Approach taken to account for treatment use	Development studies n=124 (%)	Incremental value studies n=135 (%)	Validation studies n=146 (%)
Treated patients excluded from the analysis	20 (15)	53 (39)	38 (26)
Antihypertensive medication users	4 (3)	6 (4)	6 (4)
Lipid-lowering medication users	6 (5)	10 (7)	16 (11)
Other medication users	1 (1)	2 (1)	1 (1)
Lifestyle interventions	0	0	0
Patients who received surgery	14 (10)	39 (29)	22 (15)
Untreated patients-only sensitivity analysis	9 (7)	5 (4)	4 (3)
Stratification by treatment use	1 (1)	0	0
Treatment included in the outcome	23 (19)	58 (43)	35 (24)
Treatment modelled as a predictor	54 (44)	48 (59) *	-
Antihypertensive medication use	49 (40)	44 (54) *	-
Lipid-lowering medication use	12 (10)	15 (11) *	-
Other medication use	2 (2)	5 (4) *	-
Lifestyle interventions	1 (1)	0 *	-
Surgical interventions	0	0 *	-

Table 3: Continued

Approach taken to account for treatment use	Development studies n=124 (%)	Incremental value studies n=135 (%)	Validation studies n=146 (%)
Type of treatment information modelled			
Modelled directly (not a composite predictor†)	37 (30)	44 (54) *	-
Baseline treatment	41 (33)	36 (44) *	-
Changes in treatment during follow-up	0	0 *	-
Treatment at the end of follow-up	0	1 (1) *	-
Not clearly reported	12 (10)	11 (8)	-
Statistical interactions with treatment considered	21 (17)	7 (5) *	-

* Only studies that assessed incremental value over a core set of individual predictors (n = 81) and thus had the opportunity to include treatment variables within the core set of predictors; studies that assessed incremental value over an existing prognostic model or risk score did not derive a new prediction model and are not included in the calculation. † Composite predictors are here defined as the combination of two or more variables (including treatment use) into a single predictor.

Discussion

Findings from the literature review

The use of treatments in prognostic modelling studies has not been widely addressed in cardiovascular preventative medicine. While reporting has improved over the last decade, and the majority of cardiovascular prognostic modelling studies (211 articles, 70%) made at least one reference to treatment use, we found great heterogeneity in the kinds of information and level of detail that have been reported. Only 52% of studies that developed a model reported specific information about the use of risk-lowering treatments, similar to findings from a previous review in the field of cardiovascular medicine.[6] We also confirm that information beyond baseline antihypertensive medication use, information about other treatments and changes in treatment use during follow up are frequently not reported. In addition, we found the reporting or discussion of any differences between treatment use in validation studies and their respective development studies was poorer than that observed in an earlier review of external model validation studies, which found that 40% (31/78) of articles under study discussed differences in case-mix.[14]

There are several possible explanations for the findings of the review. First, several articles used data collected during the pre-statin era,[15] which may explain why the lipid-lowering medications were scarcely reported. However, effective medications such as

aspirin and blood pressure-lowering medication have long been available, along with lifestyle interventions and some surgical procedures, which are also relevant to these studies. In addition, many articles reported a low prevalence of statin use at study baseline; in those situations it may have been assumed that treatment would not have greatly influenced the predicted probabilities. However, treatment use can greatly change over time, as shown by one study validating the AHA/ACC Pooled Cohort Equations,[16] which reported increases in antihypertensive medication use and statin use from 59.9% to 82.4% and 9.7% to 63.7% respectively over a 10-year follow-up period (1998-2007).[17] Second, while only nine articles reported that data on treatments were not available in their studies, it might be that more studies were unable to obtain such data, especially follow-up information, as this may be more costly or difficult to collect. Finally, in some studies treatments may not have been considered by the authors to be relevant to the prognostic question being addressed. One article did not model treatment effects on the grounds that "The prediction of initial CHD [coronary heart disease] events in a free-living population not on medication is emphasized",[18] i.e. the model was designed for use in individuals who are not already on treatment. However, as already discussed, this rationale does not take into account treatment drop-in that may have occurred during the follow-up period of the study.

The review is, to our knowledge, the first to give an overview of how treatment information has been reported and handled in prognostic model research. While other studies have broadly addressed related methodological issues,[14] or have focussed on a single aspect of CVD modelling, such as model development,[6] we provide comprehensive coverage of CVD prediction model studies and support this with a conceptual framework describing when and how treatments can affect a prognostic study. However, there are limitations within this study.

First, as the findings presented in the review are based on articles identified through a previously conducted systematic review, we are limited to providing information up to June 2013; more recent trends in cardiovascular prognostic modelling are not presented. Three important developments in the past four years include the ACC/AHA Pooled Cohort equations,[16] the Globorisk CVD assessment tool[19] and the QRISK3 calculator,[20] each developed as tools for the prediction of CVD in the general population. Among these three currently implemented CVD risk estimators, there is no clear consensus over how treatments should be taken into account in prognostic models for CVD; treatment use at baseline is modelled differently in each of the prognostic models, and none of the studies accounted for the effects of treatment drop-in. Thus, questions have been raised regarding the validity of these models and their respective validation studies[9,21] and treatment use remains an issue at present. Furthermore, owing to the large number of included articles (>100) published from 2009 onwards, our study provides a more up-to-date overview than previous findings.[6] As the CVD domain is a highly active field in prognostic model research, the presented results are likely optimistic for other

clinical domains; we speculate that in other clinical domains treatment use has received less attention. Second, this review focusses on a set of preventative and therapeutic treatments that modify cardiovascular risk, but may not describe all interventions that affect CVD risk. However, a detailed description is presented for the major classes of cardiovascular preventative treatments, particularly those recommended by medical guidelines. Third, as this is a review of reporting, we rely on what the authors decided to mention within the article and we cannot be entirely sure how treatment information has been collected in studies, and the extent to which it has been considered by researchers. For example, limited information could be extracted about changes in lifestyle that may have affected prognostic modelling, as this was almost never explicitly reported.

Suggestions for dealing with and reporting treatment use in prognostic model studies

Treatment use can potentially have a great impact on the reported accuracy of developed and validated prognostic models. Our review has identified that information about the use of treatments is often reported with insufficient detail to allow other researchers to evaluate the effect it may have had on the reported study findings, notably the expected predictive accuracy model in future populations. The TRIPOD statement[22,23] has already made recommendations for the reporting of information on treatment use in prognostic model studies (Item 5c), but these can be strengthened on this aspect. We provide additional recommendations for the design, analysis and reporting of prognostic model studies, to help improve the way that treatment use, in particular during follow-up, is addressed (Figure 4).

Starting with the design of future prognostic studies, we suggest that information should be collected on both treatment use at the study baseline and during follow-up, to record any changes in treatment use over time that may have impacted on the prognosis of study participants. Existing databases should contain information with enough detail to allow researchers to account for treatment use in their analyses, where necessary (see section 2.1). We provide initial recommendations on how different kinds of treatments can be taken into account when developing or validating a prediction model. This advice is based on a limited number of simulation studies, and in the absence of further simulations and empirical evidence, researchers must judge which approach will be most valid for their research. We do not provide specific guidance over how to account for complex changes in treatment use in a prognostic study, as more research is needed into the suitability of existing statistical methods. Finally, Figure 4 provides, in accordance with the TRIPOD guidelines,[23] recommendations for the minimum amount of detail that should be presented in reports of prognostic model studies. We encourage researchers to discuss the potential impact that treatment use in their study could have had on their results, including the expected accuracy of newly developed models.

Design

- Collect information on treatments used at the study baseline (see Figure 1).
- Collect information on treatment drop-in or discontinuation during follow-up (see Figure 1).
- If using readily available data (e.g. form an existing cohort or register), consider whether sufficient information on treatment use has been recorded.

Analysis

Model development

- *Guided treatments*: Consider explicitly including treatment use in the prognostic model. If a treatment was randomly allocated (e.g. data from an RCT), consider using only the subset of untreated individuals.[8]

Model validation

- *Guided treatments*: If treatments were randomly allocated, exclude treated individuals from the analysis. If treatment use is non-random (e.g. data from an observational study or register), consider first using inverse treatment probability weighting before validating the model in the untreated subset.[11]
- *Background treatments*: Consider differences in treatment use between the development and validation cohorts when exploring the impact of case-mix on model performance.[24-26]

Reporting

- Report information on treatment use at baseline. List any treatments that may have affected the prognosis of individuals in the study and the absolute number (%) treated.
- Report information on effective treatments used during follow-up and, where relevant, the duration of treatment use.
- Discuss the potential impact of treatment use on the validity and transportability of the developed prognostic model or estimates of model performance.

Figure 4: Addressing and reporting treatment use in prognostic model studies. "Treatment" refers to any medical or non-medical intervention undertaken by an individual that lowers their risk of a certain outcome.

Conclusion

In conclusion, treatment use, if ignored, can raise concerns for the transportability and validity of prognostic models. Our review shows that while the importance of treatments for prognostic prediction has been recognized in many studies, reporting rarely covers all relevant treatments, and changes in treatment have hardly been acknowledged. Furthermore, we found no clear consensus within the published literature over how treatments should be considered in the analyses of prognostic studies. Efforts should be made to collect and report detailed information about treatment use, to allow future researchers and end-users of prognostic models to more clearly identify any potential issues that treatment use may have introduced, and to understand how a model should be validated and used in practice.

Acknowledgments

The authors would like to acknowledge the following people for their contribution to the identification and selection of articles for inclusion in the original systematic review on which this review is based: James Black, Gary Collins, Thomas Debray, Pauline Heus, Lotty Hooft, Camille Lassale, Ewoud Schuit, George Siontis, René Spijker, Ioanna Tzoulaki.

References

1. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ* 2013;346:e5595.
2. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10(2):e1001381.
3. Stone NJ, Robinson JG, Lichtenstein AH, Bairey Merz CN, Blum CB, Eckel RH, et al. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014;129(25 Suppl 2):S1-45.
4. Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. NICE guidelines [CG181], National Institute for Health and Clinical Excellence 2014.
5. Perk J, De Backer G, Gohlke H, Graham I, Reiner Z, Verschuren M, et al. European Guidelines on cardiovascular disease prevention in clinical practice (version 2012). The Fifth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of nine societies and by invited experts). *Eur Heart J* 2012;33(13):1635-701.
6. Liew SM, Doust J, Glasziou P. Cardiovascular risk scores do not account for the effect of treatment: a review. *Heart* 2011;97(9):689-97.
7. Liew S, Glasziou P. Risk prediction continue to ignore treatment effects. *Br Med J Rapid Responses* 2010;340:c2442.
8. Groenwold RH, Moons KG, Pajouheshnia R, Altman DG, Collins GS, Debray TP, et al. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *J Clin Epidemiol* 2016;78:90-100.
9. Peek N, Sperrin M, Mamas M, Van Staa T, Buchan I. Hari Seldon, QRISK3, and the prediction paradox. *BMJ* 2017;357:j2099.
10. Grobbee DE, Hoes AW. *Clinical Epidemiology - Principles, Methods and Applications for Clinical Research*: London: Jones and Bartlett Publishers, 2009.
11. Pajouheshnia R, Peelen LM, Moons KGM, Reitsma JB, Groenwold RHH. Accounting for treatment use when validating a prognostic model: a simulation study. *BMC Med Res Methodol* 2017;17(1):103.
12. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;353:i2416.

13. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11(10):e1001744.
14. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;14:40.
15. Tobert JA. Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. *Nat Rev Drug Discov* 2003;2(7):517-26.
16. Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014;129(25 Suppl 2):S49-73.
17. Chia YC, Lim HM, Ching SM. Validation of the pooled cohort risk score in an Asian population - a retrospective cohort study. *BMC Cardiovasc Disord* 2014;14:163.
18. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97(18):1837-47.
19. Hajifathalian K, Ueda P, Lu Y, Woodward M, Ahmadvand A, Aguilar-Salinas CA, et al. A novel risk score to predict cardiovascular disease risk in national populations (Globorisk): a pooled analysis of prospective cohorts and health examination surveys. *Lancet Diabetes Endocrinol* 2015;3(5):339-55.
20. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;357:j2099.
21. Muntner P, Safford MM, Cushman M, Howard G. Comment on the reports of over-estimation of ASCVD risk using the 2013 AHA/ACC risk equation. *Circulation* 2014;129(2):266-7.
22. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162(1):55-63.
23. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.
24. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2014.
25. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;353:i3140.

26. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010;172(8):971-80.

Supplemental material

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #*
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	Title page
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	Abstract page
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	NR (NA- review of reporting)
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	1 (NA)
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	10
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	10
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	NR (see reference 11)
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	10
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	10-11

PRISMA 2009 Checklist Continued

Section/topic	#	Checklist item	Reported on page #*
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Additional Files 2,4
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	NA
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	NA
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	NA
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	NA
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	11-12, Figure 2
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Additional Files 2-4
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	NA
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	NA

PRISMA 2009 Checklist Continued

Section/topic	#	Checklist item	Reported on page #*
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	NA
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	NA
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	NA
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	17 (SoE= N/A)
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	18-19
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	17-20
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	20

*Values correspond to page numbers in original article *From:* Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097 For more information, visit: **www.prisma-statement.org**.

Additional file 2: List of items for data extraction

1. General study information

General study aims.

Study type.

For incremental value (IV) studies:
Is IV assessed over an existing model or a new model containing conventional predictors?

Study design.

Start of data collection.

End of data collection.

Length of follow-up.

Intended prediction horizon.

2. Reporting of treatment-specific information

Where in the article is information about treatment reported?

Is a treatment included within the definition of the outcome?
- If so, give details.

Is a treatment included within the definition of a predictor variable (composite predictor)?
- If so, give details.

Is use of any of the following treatments reported (E.g. proportion of users)?
- Cholesterol/lipid-lowering medication.
- Blood pressure-lowering/antihypertensive medication.
- Antithrombotic/anticoagulant medication.
- Lifestyle modification advice/programmes.
- Cardiovascular procedure/surgery.

If no specific details about treatment use are reported, is the collection of information about treatment use clearly reported (I.e. in the methods)?

At which stage of data collection was reported information measured (E.g. at baseline or during follow-up)?

If follow-up information is reported,
- Are incident surgical procedures reported?
- Are changes in medication use during follow-up reported?

Is treatment explicitly mentioned as part of the participant eligibility criteria?
- If so, which treatments?

Is the relevance of treatment explicitly discussed (with reference to the performance or generalizability of the model)?
- If so, provide details.

For validation studies:
Is treatment uses explicitly reported for both validation study population and the original development study population?
- If so,
  - Is there a difference in treatment use between the two sets (difference in proportion treated greater than 10%)?
  - Are the implications of any differences discussed?
    - If so, give details.

3. Accounting for treatment use in the analysis

If treatments are not accounted for in the analysis, is a reason given for why this is so?
- If so, give details.

Is the analysis restricted according to use of a treatment (I.e. Are treated individuals excluded?)?
- If so,
  - Restricted on which treatment?
  - Is restriction based on baseline status or treatment during follow-up?
  - Is this a part of a sensitivity analysis?

Is treatment modelled as a predictor?
- If so,
  - Which treatments are modelled?
  - Give details on the exact definition.
  - Is treatment modelled within a composite predictor?
  - Which kind of treatment information is modelled: baseline, follow-up, both?
  - Is treatment modelled using more advanced statistical techniques (E.g. as a time-varying covariate)?
    - If so, give details.
  - Are treatment interactions with other variables modelled?
  - Is treatment included as a predictor in the final model?
    - If not, what is the rationale behind not including the modelled treatment in the final model?
  - Is a treatment modelled alongside any associated condition (I.e. blood pressure-lowering medication and blood pressure)?

Are analyses stratified according to treatment use?

For validation studies:
Is the existing model recalibrated/updated with the specific aim of accounting for treatment use?

Additional file 3: List of articles included in the review

1. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. Circulation. 2002;106(25):3143-421. Epub 2002/12/18.
2. Aktas MK, Ozduran V, Pothier CE, Lang R, Lauer MS. Global risk scores and exercise testing for predicting all-cause mortality in a preventive medicine program. JAMA. 2004;292(12):1462-8.
3. Alssema M, Newson RS, Bakker SJL, Stehouwer CDA, Heymans MW, Nijpels G, et al. One risk assessment tool for cardiovascular disease, type 2 diabetes, and chronic kidney disease. Diabetes Care. 2012;35(4):741-8.
4. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. Am Heart J. 1991;121(1 Pt 2):293-8. Epub 1991/01/01.
5. Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk profile. A statement for health professionals. Circulation. 1991;83(1):356-62. Epub 1991/01/01.
6. Araujo AB, Hall SA, Ganz P, Chiu GR, Rosen RC, Kupelian V, et al. Does erectile dysfunction contribute to cardiovascular disease risk prediction beyond the Framingham risk score? J Am Coll Cardiol. 2010;55(4):350-6.
7. Arima H, Yonemoto K, Doi Y, Ninomiya T, Hata J, Tanizaki Y, et al. Development and validation of a cardiovascular risk prediction model for Japanese: the Hisayama study. Hypertens Res. 2009;32(12):1119-22.
8. Asayama K, Ohkubo T, Sato A, Hara A, Obara T, Yasui D, et al. Proposal of a risk-stratification system for the Japanese population based on blood pressure levels: the Ohasama study. Hypertens Res. 2008;31(7):1315-22. Epub 2008/10/30.
9. Asia Pacific Cohort Studies Collaboration. Coronary risk prediction for those with and without diabetes. Eur J Cardiovasc Prev Rehabil. 2006;13(1):30-6. Epub 2006/02/02.
10. Asia Pacific Cohort Studies Collaboration, Barzi F, Patel A, Gu D, Sritara P, Lam TH, et al. Cardiovascular risk prediction tools for populations in Asia. J Epidemiol Community Health. 2007;61(2):115-21.
11. Aslibekyan S, Campos H, Loucks EB, Linkletter CD, Ordovas JM, Baylin A. Development of a cardiovascular risk score for use in low- and middle-income countries. J Nutr. 2011;141(7):1375-80.
12. Asselbergs FW, Hillege HL, van Gilst WH. Framingham score and microalbuminuria: combined future targets for primary prevention? Kidney Int Suppl. 2004(92):S111-4.
13. Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. Circulation. 2002;105(3):310-5. Epub 2002/01/24.
14. Assmann G, Schulte H, Cullen P, Seedorf U. Assessing risk of myocardial infarction and stroke: new data from the Prospective Cardiovascular Munster (PROCAM) study. Eur J Clin Invest. 2007;37(12):925-32.

15. Assmann G, Schulte H, Seedorf U. Cardiovascular risk assessment in the metabolic syndrome: results from the Prospective Cardiovascular Munster (PROCAM) Study. Int J Obes. 2008;32 Suppl 2:S11-6.
16. Badheka AO, Patel N, Tuliani TA, Rathod A, Marzouka GR, Zalawadiya S, et al. Electrocardiographic abnormalities and reclassification of cardiovascular risk: insights from NHANES-III. Am J Med. 2013;126(4):319-26.e2.
17. Baik I, Cho NH, Kim SH, Shin C. Dietary information improves cardiovascular disease risk prediction models. Eur J Clin Nutr. 2013;67(1):25-30.
18. Baldassarre D, Hamsten A, Veglia F, de Faire U, Humphries SE, Smit AJ, et al. Measurements of carotid intima-media thickness and of interadventitia common carotid diameter improve prediction of cardiovascular events: results of the IMPROVE (Carotid Intima Media Thickness [IMT] and IMT-Progression as Predictors of Vascular Events in a High Risk European Population) study. J Am Coll Cardiol. 2012;60(16):1489-99.
19. Balkau B, Hu G, Qiao Q, Tuomilehto J, Borch-Johnsen K, Pyorala K, et al. Prediction of the risk of cardiovascular mortality using a score that includes glucose as a risk factor. The DECODE Study. Diabetologia. 2004;47(12):2118-28.
20. Bare LA, Morrison AC, Rowland CM, Shiffman D, Luke MM, Iakoubova OA, et al. Five common gene variants identify elevated genetic risk for coronary heart disease. Genet Med. 2007;9(10):682-9.
21. Barroso LC, Muro EC, Herrera ND, Ochoa GF, Hueros JIC, Buitrago F. Performance of the Framingham and SCORE cardiovascular risk prediction functions in a non-diabetic population of a Spanish health care centre: a validation study. Scand J Prim Health Care. 2010;28(4):242-8.
22. Bastuji-Garin S, Deverly A, Moyse D, Castaigne A, Mancia G, de Leeuw PW, et al. The Framingham prediction rule is not valid in a European population of treated hypertensive patients. J Hypertens. 2002;20(10):1973-80.
23. Baxi NS, Jackson JL, Ritter J, Sessums LL. How well do the Framingham risk factors correlate with diagnoses of ischemic heart disease and cerebrovascular disease in a military beneficiary cohort? Mil Med. 2011;176(4):408-13.
24. Becker CR, Majeed A, Crispin A, Knez A, Schoepf UJ, Boekstegers P, et al. CT measurement of coronary calcium mass: impact on global cardiac risk assessment. Eur Radiol. 2005;15(1):96-101.
25. Beer C, Alfonso H, Flicker L, Norman PE, Hankey GJ, Almeida OP. Traditional risk factors for incident cardiovascular events have limited importance in later life compared with the health in men study cardiovascular risk score. Stroke. 2011;42(4):952-9.
26. Bell K, Hayen A, McGeechan K, Neal B, Irwig L. Effects of additional blood pressure and lipid measurements on the prediction of cardiovascular risk. Eur J Prev Cardiol. 2012;19(6):1474-85.
27. Berard E, Bongard V, Arveiler D, Amouyel P, Wagner A, Dallongeville J, et al. Ten-year risk of all-cause mortality: assessment of a risk prediction algorithm in a French general population. Eur J Epidemiol. 2011;26(5):359-68.
28. Berry JD, Lloyd-Jones DM, Garside DB, Greenland P. Framingham risk score and prediction of coronary heart disease death in young men. Am Heart J. 2007;154(1):80-6.

29. Bhopal R, Fischbacher C, Vartiainen E, Unwin N, White M, Alberti G. Predicted and observed cardiovascular disease in South Asians: application of FINRISK, Framingham and SCORE models to Newcastle Heart Project data. J Public Health. 2005;27(1):93-100.
30. Bineau S, Dufouil C, Helmer C, Ritchie K, Empana J-P, Ducimetiere P, et al. Framingham stroke risk function in a large population-based cohort of elderly people: the 3C study. Stroke. 2009;40(5):1564-70.
31. Boland B, De Muylder R, Goderis G, Degryse J, Gueuning Y, Paulus D, et al. Cardiovascular prevention in general practice: development and validation of an algorithm. Acta Cardiol. 2004;59(6):598-605.
32. Bolton JL, Stewart MCW, Wilson JF, Anderson N, Price JF. Improvement in Prediction of Coronary Heart Disease Risk over Conventional Risk Factors Using SNPs Identified in Genome-Wide Association Studies. PLoS ONE. 2013;8(2).
33. Boudik F, Reissigova J, Hrach K, Tomeckova M, Bultas J, Anger Z, et al. Primary prevention of coronary artery disease among middle aged men in Prague: twenty-year follow-up results. Atherosclerosis. 2006;184(1):86-93. Epub 2005/11/19.
34. Boyar A. Creating a web application that combines Framingham risk with Electron Beam CT Coronary Calcium Score to calculate a new event risk. J Thorac Imaging. 2006;21(1):91-6.
35. Bozorgmanesh M, Hadaegh F, Azizi F. Predictive accuracy of the 'Framingham's general CVD algorithm' in a Middle Eastern population: Tehran Lipid and Glucose Study. Int J Clin Pract. 2011;65(3):264-73.
36. Brand RJ, Rosenman RH, Sholtz RI, Friedman M. Multivariate prediction of coronary heart disease in the Western Collaborative Group Study compared to the findings of the Framingham study. Circulation. 1976;53(2):348-55. Epub 1976/02/01.
37. Braun J, Bopp M, Faeh D. Blood glucose may be an alternative to cholesterol in CVD risk prediction charts. Cardiovasc Diabetol. 2013;12(1).
38. Brautbar A, Ballantyne CM, Lawson K, Nambi V, Chambless L, Folsom AR, et al. Impact of adding a single allele in the 9p21 locus to traditional risk factors on reclassification of coronary heart disease risk and implications for lipid-modifying therapy in the Atherosclerosis Risk in Communities study. Circ Cardiovasc Genet. 2009;2(3):279-85.
39. Brautbar A, Pompeii LA, Dehghan A, Ngwa JS, Nambi V, Virani SS, et al. A genetic risk score based on direct associations with coronary heart disease improves coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC), but not in the Rotterdam and Framingham Offspring, Studies. Atherosclerosis. 2012;223(2):421-6.
40. Brindle P, Emberson J, Lampe F, Walker M, Whincup P, Fahey T, et al. Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study. BMJ. 2003;327(7426):1267. Epub 2003/12/04.
41. Brindle P, May M, Gill P, Cappuccio F, D'Agostino R, Sr., Fischbacher C, et al. Primary prevention of cardiovascular disease: a web-based risk score for seven British black and minority ethnic groups. Heart. 2006;92(11):1595-602. Epub 2006/06/10.

42. Brindle PM, McConnachie A, Upton MN, Hart CL, Davey Smith G, Watt GCM. The accuracy of the Framingham risk-score in different socioeconomic groups: a prospective study. Br J Gen Pract. 2005;55(520):838-45.
43. Brunner EJ, Shipley MJ, Marmot MG, Kivimaki M, Witte DR. Do the Joint British Society (JBS2) guidelines on prevention of cardiovascular disease with respect to plasma glucose improve risk stratification in the general population? Prospective cohort study. Diabet Med. 2010;27(5):550-5.
44. Buitrago F, Calvo-Hueros JI, Canon-Barroso L, Pozuelos-Estrada G, Molina-Martinez L, Espigares-Arroyo M, et al. Original and REGICOR Framingham functions in a nondiabetic population of a Spanish health care center: a validation study. Ann Fam Med. 2011;9(5):431-8.
45. Canoui-Poitrine F, Luc G, Mallat Z, Machez E, Bingham A, Ferrieres J, et al. Systemic chemokine levels, coronary heart disease, and ischemic stroke events: the PRIME study. Neurology. 2011;77(12):1165-73.
46. Cao JJ, Arnold AM, Manolio TA, Polak JF, Psaty BM, Hirsch CH, et al. Association of carotid artery intima-media thickness, plaques, and C-reactive protein with future cardiovascular disease and all-cause mortality: The cardiovascular health study. Circulation. 2007;116(1):32-8.
47. Chamberlain AM, Agarwal SK, Folsom AR, Soliman EZ, Chambless LE, Crow R, et al. A clinical risk score for atrial fibrillation in a biracial prospective cohort (from the Atherosclerosis Risk in Communities [ARIC] study). Am J Cardiol. 2011;107(1):85-91.
48. Chambless LE, Folsom AR, Sharrett AR, Sorlie P, Couper D, Szklo M, et al. Coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC) study. J Clin Epidemiol. 2003;56(9):880-90. Epub 2003/09/25.
49. Chambless LE, Heiss G, Shahar E, Earp MJ, Toole J. Prediction of ischemic stroke risk in the Atherosclerosis Risk in Communities Study.[Erratum appears in Am J Epidemiol. 2004 Nov 1;160(9):927]. Am J Epidemiol. 2004;160(3):259-69.
50. Chamnan P, Simmons RK, Hori H, Sharp S, Khaw K-T, Wareham NJ, et al. A simple risk score using routine data for predicting cardiovascular disease in primary care. Br J Gen Pract. 2010;60(577):e327-34.
51. Chen L, Tonkin AM, Moon L, Mitchell P, Dobson A, Giles G, et al. Recalibration and validation of the SCORE risk chart in the Australian population: the AusSCORE chart. Eur J Cardiovasc Prev Rehabil. 2009;16(5):562-70.
52. Chien KL, Hsu HC, Su TC, Chang WT, Chen PC, Sung FC, et al. Constructing a point-based prediction model for the risk of coronary artery disease in a Chinese community: A report from a cohort study in Taiwan. Int J Cardiol. 2012;157(2):263-8.
53. Chien KL, Su TC, Hsu HC, Chang WT, Chen PC, Sung FC, et al. Constructing the prediction model for the risk of stroke in a Chinese population: report from a cohort study in Taiwan. Stroke. 2010;41(9):1858-64. Epub 2010/07/31.
54. Chironi G, Simon A, Megnien J-L, Sirieix M-E, Mousseaux E, Pessana F, et al. Impact of coronary artery calcium on cardiovascular risk categorization and lipid-lowering drug eligibility in asymptomatic hypercholesterolemic men. Int J Cardiol. 2011;151(2):200-4.

55. Church TS, Levine BD, McGuire DK, Lamonte MJ, Fitzgerald SJ, Cheng YJ, et al. Coronary artery calcium score, risk factors, and incident coronary heart disease events. Atherosclerosis. 2007;190(1):224-31.
56. Ciampi A, Courteau J, Niyonsenga T, Xhignesse M, Lussier-Cacan S, Roy M. Family history and the risk of coronary heart disease: comparing predictive models. Eur J Epidemiol. 2001;17(7):609-20. Epub 2002/06/28.
57. Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. BMJ. 2009;339:b2584.
58. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. BMJ. 2010;340:c2442.
59. Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. BMJ. 2012;344:e4181. Epub 2012/06/23.
60. Comin E, Solanas P, Cabezas C, Subirana I, Ramos R, Gene-Badia J, et al. Estimating cardiovascular risk in Spain using different algorithms. Rev Esp Cardiol. 2007;60(7):693-702.
61. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003;24(11):987-1003. Epub 2003/06/06.
62. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. Ann Intern Med. 2006;145(1):21-9.
63. Cook NR, Paynter NP, Eaton CB, Manson JE, Martin LW, Robinson JG, et al. Comparison of the Framingham and Reynolds Risk scores for global cardiovascular risk prediction in the multiethnic Women's Health Initiative. Circulation. 2012;125(14):1748-56, S1-11.
64. Cooney MT, Dudina A, De Bacquer D, Fitzgerald A, Conroy R, Sans S, et al. How much does HDL cholesterol add to risk estimation? A report from the SCORE Investigators. Eur J Cardiovasc Prev Rehabil. 2009;16(3):304-14.
65. Cooney MT, Vartiainen E, Laatikainen T, Joulevi A, Dudina A, Graham I. Simplifying cardiovascular risk estimation using resting heart rate. Eur Heart J. 2010;31(17):2141-7.
66. Cooper JA, Miller GJ, Humphries SE. A comparison of the PROCAM and Framingham point-scoring systems for estimation of individual risk of coronary heart disease in the Second Northwick Park Heart Study. Atherosclerosis. 2005;181(1):93-100.
67. Cournot M, Bura A, Cambou J-P, Taraszkiewicz D, Maloizel J, Galinier M, et al. Arterial ultrasound screening as a tool for coronary risk assessment in asymptomatic men and women. Angiology. 2012;63(4):282-8.
68. Cournot M, Taraszkiewicz D, Cambou J-P, Galinier M, Boccalon H, Hanaire-Broutin H, et al. Additional prognostic value of physical examination, exercise testing, and arterial ultrasonography for coronary risk assessment in primary prevention. Am Heart J. 2009;158(5):845-51.

69. Cournot M, Taraszkiewicz D, Galinier M, Chamontin B, Boccalon H, Hanaire-Broutin H, et al. Is exercise testing useful to improve the prediction of coronary events in asymptomatic subjects? Eur J Cardiovasc Prev Rehabil. 2006;13(1):37-44.
70. Cross DS, McCarty CA, Hytopoulos E, Beggs M, Nolan N, Harrington DS, et al. Coronary risk assessment among intermediate risk patients using a clinical and biomarker based algorithm developed and validated in two population cohorts. Curr Med Res Opin. 2012;28(11):1819-30.
71. Cushman M, Arnold AM, Psaty BM, Manolio TA, Kuller LH, Burke GL, et al. C-reactive protein and the 10-year incidence of coronary heart disease in older men and women: the cardiovascular health study. Circulation. 2005;112(1):25-31.
72. D'Agostino RB, Sr., Grundy S, Sullivan LM, Wilson P. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. JAMA. 2001;286(2):180-7. Epub 2001/07/13.
73. D'Agostino RB, Russell MW, Huse DM, Ellison RC, Silbershatz H, Wilson PW, et al. Primary and subsequent coronary risk appraisal: new results from the Framingham study. Am Heart J. 2000;139(2 Pt 1):272-81. Epub 2000/01/29.
74. D'Agostino RB, Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117(6):743-53. Epub 2008/01/24.
75. D'Agostino RB, Wolf PA, Belanger AJ, Kannel WB. Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study. Stroke. 1994;25(1):40-3. Epub 1994/01/01.
76. Davies RW, Dandona S, Stewart AFR, Chen L, Ellis SG, Tang WHW, et al. Improved prediction of cardiovascular disease based on a panel of single nucleotide polymorphisms identified through genome-wide association studies. Circ Cardiovasc Genet. 2010;3(5):468-74.
77. De Bacquer D, De Backer G. Predictive ability of the SCORE Belgium risk chart for cardiovascular mortality. Int J Cardiol. 2010;143(3):385-90.
78. de la Iglesia B, Potter JF, Poulter NR, Robins MM, Skinner J. Performance of the ASSIGN cardiovascular disease risk score on a UK cohort of patients from general practice. Heart. 2011;97(6):491-9.
79. de Ruijter W, Westendorp RGJ, Assendelft WJJ, den Elzen WPJ, de Craen AJM, le Cessie S, et al. Use of Framingham risk score and new biomarkers to predict cardiovascular mortality in older people: population based observational cohort study. BMJ. 2009;338:a3083.
80. DECODE Study Group. Does diagnosis of the metabolic syndrome detect further men at high risk of cardiovascular death beyond those identified by a conventional cardiovascular risk score? The DECODE Study. Eur J Cardiovasc Prev Rehabil. 2007;14(2):192-9. Epub 2007/04/21.
81. Denes P, Larson JC, Lloyd-Jones DM, Prineas RJ, Greenland P. Major and minor ECG abnormalities in asymptomatic women and risk of cardiovascular events and mortality. JAMA. 2007;297(9):978-85.
82. Detrano R, Guerci AD, Carr JJ, Bild DE, Burke G, Folsom AR, et al. Coronary calcium as a predictor of coronary events in four racial or ethnic groups. N Engl J Med. 2008;358(13):1336-45.

83. Dhamoon MS, Moon YP, Paik MC, Sacco RL, Elkind MSV. The inclusion of stroke in risk stratification for primary prevention of vascular events: the Northern Manhattan Study. Stroke. 2011;42(10):2878-82.
84. Ding K, Bailey KR, Kullo IJ. Genotype-informed estimation of risk of coronary heart disease based on genome-wide association data linked to the electronic medical record. BMC Cardiovasc Disord. 2011;11:66.
85. Diverse Populations Collaborative Group. Prediction of mortality from coronary heart disease among diverse populations: is there a common predictive function? Heart. 2002;88(3):222-8. Epub 2002/08/16.
86. Donfrancesco C, Palmieri L, Cooney M-T, Vanuzzo D, Panico S, Cesana G, et al. Italian cardiovascular mortality charts of the CUORE project: are they comparable with the SCORE charts? Eur J Cardiovasc Prev Rehabil. 2010;17(4):403-9.
87. Drawz PE, Baraniuk S, Davis BR, Brown CD, Colon PJ, Sr., Cujyet AB, et al. Cardiovascular risk assessment: addition of CKD and race to the Framingham equation. Am Heart J. 2012;164(6):925-31.e2.
88. Dunder K, Lind L, Zethelius B, Berglund L, Lithell H. Evaluation of a scoring scheme, including proinsulin and the apolipoprotein B/apolipoprotein A1 ratio, for the risk of acute coronary events in middle-aged men: Uppsala Longitudinal Study of Adult Men (ULSAM). Am Heart J. 2004;148(4):596-601.
89. Duprez DA, Florea N, Zhong W, Grandits GA, Hawthorne CK, Hoke L, et al. Vascular and cardiac functional and structural screening to identify risk of future morbid events: preliminary observations. J Am Soc Hypertens. 2011;5(5):401-9. Epub 2011/07/02.
90. Dutta A, Henley W, Lang IA, Murray A, Guralnik J, Wallace RB, et al. The coronary artery disease-associated 9p21 variant and later life 20-year survival to cohort extinction. Circulation Cardiovascular Genetics. 2011;4(5):542-8.
91. Dutta A, Henley W, Pilling LC, Wallace RB, Melzer D. Uric acid measurement improves prediction of cardiovascular mortality in later life. J Am Geriatr Soc. 2013;61(3):319-26.
92. Emerging Risk Factors Collaboration, Di Angelantonio E, Gao P, Pennells L, Kaptoge S, Caslake M, et al. Lipid-related markers and cardiovascular disease prediction. JAMA. 2012;307(23):2499-506.
93. Empana JP, Ducimetiere P, Arveiler D, Ferrieres J, Evans A, Ruidavets JB, et al. Are the Framingham and PROCAM coronary heart disease risk functions applicable to different European populations? The PRIME Study. Eur Heart J. 2003;24(21):1903-11.
94. Empana JP, Tafflet M, Escolano S, Vergnaux AC, Bineau S, Ruidavets JB, et al. Predicting CHD risk in France: A pooled analysis of the D.E.S.I.R., Three City, PRIME, and SU.VI.MAX studies. Eur J Cardiovasc Prev Rehabil. 2011;18(2):175-85.
95. Erbel R, Mohlenkamp S, Lehmann N, Schmermund A, Moebus S, Stang A, et al. Sex related cardiovascular risk stratification based on quantification of atherosclerosis and inflammation. Atherosclerosis. 2008;197(2):662-72. Epub 2007/03/28.

96. Erbel R, Mohlenkamp S, Moebus S, Schmermund A, Lehmann N, Stang A, et al. Coronary risk stratification, discrimination, and reclassification improvement based on quantification of subclinical coronary atherosclerosis: the Heinz Nixdorf Recall study. J Am Coll Cardiol. 2010;56(17):1397-406.
97. Erikssen G, Bodegard J, Bjornholt JV, Liestol K, Thelle DS, Erikssen J. Exercise testing of healthy men in a new perspective: from diagnosis to prognosis. Eur Heart J. 2004;25(11):978-86.
98. Faeh D, Braun J, Rufibach K, Puhan MA, Marques-Vidal P, Bopp M. Population Specific and Up to Date Cardiovascular Risk Charts Can Be Efficiently Obtained with Record Linkage of Routine and Observational Data. PLoS ONE. 2013;8(2).
99. Ferrario M, Chiodini P, Chambless LE, Cesana G, Vanuzzo D, Panico S, et al. Prediction of coronary events in a low incidence population. Assessing accuracy of the CUORE Cohort Study prediction equation. Int J Epidemiol. 2005;34(2):413-21.
100. Fiscella K, Tancredi D, Franks P. Adding socioeconomic status to Framingham scoring to reduce disparities in coronary risk assessment. Am Heart J. 2009;157(6):988-94.
101. Folsom AR, Chambless LE, Duncan BB, Gilbert AC, Pankow JS, Atherosclerosis Risk in Communities Study I. Prediction of coronary heart disease in middle-aged adults with diabetes. Diabetes Care. 2003;26(10):2777-84.
102. Franks P, Tancredi DJ, Winters P, Fiscella K. Including socioeconomic status in coronary heart disease risk estimation. Ann Fam Med. 2010;8(5):447-53.
103. Friedland DR, Cederberg C, Tarima S. Audiometric pattern as a predictor of cardiovascular status: development of a model for assessment of risk. Laryngoscope. 2009;119(3):473-86.
104. Gaziano TA, Young CR, Fitzmaurice G, Atwood S, Gaziano JM. Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the NHANES I Follow-up Study cohort. Lancet. 2008;371(9616):923-31.
105. Glynn RJ, L'Italien GJ, Sesso HD, Jackson EA, Buring JE. Development of predictive models for long-term cardiovascular risk associated with systolic and diastolic blood pressure. Hypertension. 2002;39(1):105-10. Epub 2002/01/19.
106. Greenland P, LaBree L, Azen SP, Doherty TM, Detrano RC. Coronary artery calcium score combined with Framingham score for risk prediction in asymptomatic individuals.[Erratum appears in JAMA. 2004 Feb 4;291(5):563]. JAMA. 2004;291(2):210-5.
107. Gulati M, Arnsdorf MF, Shaw LJ, Pandey DK, Thisted RA, Lauderdale DS, et al. Prognostic value of the duke treadmill score in asymptomatic women. Am J Cardiol. 2005;96(3):369-75.
108. Hadaegh F, Mohebi R, Bozorgmanesh M, Saadat N, Sheikholeslami F, Azizi F. Electrocardiographic abnormalities improve classification of coronary heart disease risk in women: Tehran Lipid and Glucose Study. Atherosclerosis. 2012;222(1):110-5.

109. Haluska BA, Jeffries L, Carlier S, Marwick TH. Measurement of arterial distensibility and compliance to assess prognosis. Atherosclerosis. 2010;209(2):474-80.
110. Hamer M, Chida Y, Stamatakis E. Utility of C-reactive protein for cardiovascular risk stratification across three age groups in subjects without existing cardiovascular diseases. Am J Cardiol. 2009;104(4):538-42.
111. Hense HW, Schulte H, Lowel H, Assmann G, Keil U. Framingham risk function overestimates risk of coronary heart disease in men and women from Germany--results from the MONICA Augsburg and the PROCAM cohorts. Eur Heart J. 2003;24(10):937-45. Epub 2003/04/26.
112. Hense H-W, Koesters E, Wellmann J, Meisinger C, Volzke H, Keil U. Evaluation of a recalibrated Systematic Coronary Risk Evaluation cardiovascular risk chart: results from Systematic Coronary Risk Evaluation Germany. Eur J Cardiovasc Prev Rehabil. 2008;15(4):409-15.
113. Hippisley-Cox J, Coupland C, Robson J, Brindle P. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. BMJ. 2010;341:c6624.
114. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. Heart. 2008;94(1):34-9.
115. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ. 2007;335(7611):136.
116. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008;336(7659):1475-82.
117. Hoes AW, Grobbee DE, Valkenburg HA, Lubsen J, Hofman A. Cardiovascular risk and all-cause mortality; a 12 year follow-up study in The Netherlands. Eur J Epidemiol. 1993;9(3):285-92. Epub 1993/05/01.
118. Houterman S, Boshuizen HC, Verschuren WM, Giampaoli S, Nissinen A, Menotti A, et al. Predicting cardiovascular risk in the elderly in different European countries. Eur Heart J. 2002;23(4):294-300. Epub 2002/01/29.
119. Hsia J, Rodabough RJ, Manson JE, Liu S, Freiberg MS, Graettinger W, et al. Evaluation of the American Heart Association cardiovascular disease prevention guideline for women. Circ Cardiovasc Qual Outcomes. 2010;3(2):128-34.
120. Hughes MF, Saarela O, Blankenberg S, Zeller T, Havulinna AS, Kuulasmaa K, et al. A multiple biomarker risk score for guiding clinical decisions using a decision curve approach. Eur J Prev Cardiol. 2012;19(4):874-84.
121. Hughes MF, Saarela O, Stritzke J, Kee F, Silander K, Klopp N, et al. Genetic markers enhance coronary risk prediction in men: the MORGAM prospective cohorts. PLoS ONE. 2012;7(7):e40922.
122. Humphries SE, Cooper JA, Talmud PJ, Miller GJ. Candidate gene genotypes, along with conventional risk factor assessment, improve estimation of coronary heart disease risk in healthy UK men. Clin Chem. 2007;53(1):8-16.

123. Hurley LP, Dickinson LM, Estacio RO, Steiner JF, Havranek EP. Prediction of cardiovascular death in racial/ethnic minorities using Framingham risk factors. Circ Cardiovasc Qual Outcomes. 2010;3(2):181-7.
124. Iqbal FM, Al Jaroudi W, Sanam K, Sweeney A, Heo J, Iskandrian AE, et al. Reclassification of cardiovascular risk in patients with normal myocardial perfusion imaging using heart rate response to vasodilator stress. Am J Cardiol. 2013;111(2):190-5.
125. Ishikawa S, Matsumoto M, Kayaba K, Gotoh T, Nago N, Tsutsumi A, et al. Risk charts illustrating the 10-year risk of stroke among residents of Japanese rural communities: the JMS Cohort Study. J Epidemiol. 2009;19(2):101-6.
126. Ito H, Pacold IV, Durazo-Arvizu R, Liu K, Shilipak MG, Goff DC, Jr., et al. The effect of including cystatin C or creatinine in a cardiovascular risk model for asymptomatic individuals: the multi-ethnic study of atherosclerosis. Am J Epidemiol. 2011;174(8):949-57.
127. Jalal D, Chonchol M, Etgen T, Sander D. C-reactive protein as a predictor of cardiovascular events in elderly patients with chronic kidney disease. J Nephrol. 2012;25(5):719-25.
128. Janssen I, Katzmarzyk PT, Church TS, Blair SN. The Cooper Clinic Mortality Risk Index: clinical score sheet for men. Am J Prev Med. 2005;29(3):194-203.
129. Jimenez-Corona A, Lopez-Ridaura R, Williams K, Gonzalez-Villalpando ME, Simon J, Gonzalez-Villalpando C. Applicability of Framingham risk equations for studying a low-income Mexican population. Salud Publica Mex. 2009;51(4):298-305.
130. Kamstrup PR, Tybjaerg-Hansen A, Nordestgaard BG. Extreme lipoprotein(a) levels and improved cardiovascular risk prediction. J Am Coll Cardiol. 2013;61(11):1146-56.
131. Kang HM, Kim D-J. Metabolic Syndrome versus Framingham Risk Score for Association of Self-Reported Coronary Heart Disease: The 2005 Korean Health and Nutrition Examination Survey. Diabetes Metab J. 2012;36(3):237-44.
132. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. Am J Cardiol. 1976;38(1):46-51. Epub 1976/07/01.
133. Kathiresan S, Melander O, Anevski D, Guiducci C, Burtt NP, Roos C, et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. N Engl J Med. 2008;358(12):1240-9.
134. Katz D, Foxman B. How well do prediction equations predict? Using receiver operating characteristic curves and accuracy curves to compare validity and generalizability. Epidemiology. 1993;4(4):319-26. Epub 1993/07/01.
135. Ketola E, Laatikainen T, Vartiainen E. Evaluating risk for cardiovascular diseases--vain or value? How do different cardiovascular risk scores act in real life. Eur J Public Health. 2010;20(1):107-12.
136. Keys A, Aravanis C, Blackburn H, Van Buchem FS, Buzina R, Djordjevic BS, et al. Probability of middle-aged men developing coronary heart disease in five years. Circulation. 1972;45(4):815-28. Epub 1972/04/01.
137. Khalili D, Hadaegh F, Soori H, Steyerberg EW, Bozorgmanesh M, Azizi F. Clinical usefulness of the Framingham cardiovascular risk profile beyond its statistical performance: the Tehran Lipid and Glucose Study. Am J Epidemiol. 2012;176(3):177-86.

138. Knuiman MW, Vu HT. Prediction of coronary heart disease mortality in Busselton, Western Australia: an evaluation of the Framingham, national health epidemiologic follow up study, and WHO ERICA risk scores. J Epidemiol Community Health. 1997;51(5):515-9. Epub 1998/01/13.
139. Knuiman MW, Vu HT, Bartholomew HC. Multivariate risk estimation for coronary heart disease: the Busselton Health Study. Aust N Z J Public Health. 1998;22(7):747-53. Epub 1999/01/16.
140. Koizumi J, Shimizu M, Miyamoto S, Takeda R, Ohka T, Kanaya H, et al. Risk evaluation of coronary heart disease and cerebrovascular disease by the Japan Atherosclerosis Society Guidelines 2002 using the cohort of the Holicos-PAT study. J Atheroscler Thromb. 2005;12(1):48-52.
141. Koller MT, Leening MJG, Wolbers M, Steyerberg EW, Hunink MGM, Schoop R, et al. Development and validation of a coronary risk prediction model for older U.S. and European persons in the cardiovascular health study and the Rotterdam Study. Ann Intern Med. 2012;157(6):389-97.
142. Koller MT, Steyerberg EW, Wolbers M, Stijnen T, Bucher HC, Hunink MGM, et al. Validity of the Framingham point scores in the elderly: results from the Rotterdam study. Am Heart J. 2007;154(1):87-93.
143. Larson MG. Assessment of cardiovascular risk factors in the elderly: the Framingham Heart Study. Stat Med. 1995;14(16):1745-56. Epub 1995/08/30.
144. Laurier D, Nguyen PC, Cazelles B, Segond P. Estimation of CHD risk in a French working population using a modified Framingham model. The PCV-METRA Group. J Clin Epidemiol. 1994;47(12):1353-64. Epub 1994/12/01.
145. Leaverton PE, Sorlie PD, Kleinman JC, Dannenberg AL, Ingster-Moore L, Kannel WB, et al. Representativeness of the Framingham risk model for coronary heart disease mortality: a comparison with a national cohort study. J Chronic Dis. 1987;40(8):775-84. Epub 1987/01/01.
146. Lee ET, Howard BV, Wang W, Welty TK, Galloway JM, Best LG, et al. Prediction of coronary heart disease in a population with high prevalence of diabetes and albuminuria: the Strong Heart Study. Circulation. 2006;113(25):2897-905.
147. Lee J, Heng D, Ma S, Chew S-K, Hughes K, Tai ES. The metabolic syndrome and mortality: the Singapore Cardiovascular Cohort Study. Clin Endocrinol (Oxf). 2008;69(2):225-30.
148. Levy D, Wilson PW, Anderson KM, Castelli WP. Stratifying the patient at risk from coronary disease: new insights from the Framingham Heart Study. Am Heart J. 1990;119(3 Pt 2):712-7; discussion 7. Epub 1990/03/01.
149. Lindman AS, Veierod MB, Pedersen JI, Tverdal A, Njolstad I, Selmer R. The ability of the SCORE high-risk model to predict 10-year cardiovascular disease mortality in Norway. Eur J Cardiovasc Prev Rehabil. 2007;14(4):501-7.
150. L'Italien G, Ford I, Norrie J, LaPuerta P, Ehreth J, Jackson J, et al. The cardiovascular event reduction tool (CERT)--a simplified cardiac risk prediction model developed from the West of Scotland Coronary Prevention Study (WOSCOPS). Am J Cardiol. 2000;85(6):720-4. Epub 2002/05/10.
151. Liu J, Hong Y, D'Agostino RB, Sr., Wu Z, Wang W, Sun J, et al. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study. JAMA. 2004;291(21):2591-9.

152. Lloyd-Jones DM, Wilson PWF, Larson MG, Beiser A, Leip EP, D'Agostino RB, et al. Framingham risk score and prediction of lifetime risk for coronary heart disease. Am J Cardiol. 2004;94(1):20-4.
153. Lumley T, Kronmal RA, Cushman M, Manolio TA, Goldstein S. A stroke prediction score in the elderly: validation and Web-based application. J Clin Epidemiol. 2002;55(2):129-36. Epub 2002/01/26.
154. Macfarlane PW, Norrie J. The value of the electrocardiogram in risk assessment in primary prevention: Experience from the West of Scotland Coronary Prevention Study. J Electrocardiol. 2007;40(1):101-9.
155. Mainous AG, 3rd, Everett CJ, Player MS, King DE, Diaz VA. Importance of a patient's personal health history on assessments of future risk of coronary heart disease. J Am Board Fam Med. 2008;21(5):408-13.
156. Mainous AG, 3rd, Koopman RJ, Diaz VA, Everett CJ, Wilson PWF, Tilley BC. A coronary heart disease risk score based on patient-reported information. Am J Cardiol. 2007;99(9):1236-41.
157. Manickam P, Rathod A, Panaich S, Hari P, Veeranna V, Badheka A, et al. Comparative prognostic utility of conventional and novel lipid parameters for cardiovascular disease risk prediction: do novel lipid parameters offer an advantage? J Clin Lipidol. 2011;5(2):82-90.
158. Mannan H, Stevenson C, Peeters A, Walls H, McNeil J. Framingham risk prediction equations for incidence of cardiovascular disease using detailed measures for smoking. Heart Int. 2010;5(2):e11.
159. Mannan HR, Stevenson CE, Peeters A, McNeil JJ. A new set of risk equations for predicting long term risk of all-cause mortality using cardiovascular risk factors. Prev Med. 2013;56(1):41-5.
160. Mannan HR, Stevenson CE, Peeters A, Walls HL, McNeil JJ. Age at quitting smoking as a predictor of risk of cardiovascular disease incidence independent of smoking status, time since quitting and pack-years. BMC Research Notes. 2011;4:39.
161. Marrugat J, D'Agostino R, Sullivan L, Elosua R, Wilson P, Ordovas J, et al. An adaptation of the Framingham coronary heart disease risk function to European Mediterranean areas. J Epidemiol Community Health. 2003;57(8):634-8. Epub 2003/07/29.
162. Marrugat J, Solanas P, D'Agostino R, Sullivan L, Ordovas J, Cordon F, et al. Coronary risk estimation in Spain using a calibrated Framingham function. Rev Esp Cardiol. 2003;56(3):253-61. Epub 2003/03/08. Estimacion del riesgo coronario en Espana mediante la ecuacion de Framingham calibrada.
163. Marrugat J, Subirana I, Comin E, Cabezas C, Vila J, Elosua R, et al. Validity of an adaptation of the Framingham cardiovascular risk function: The VERIFICA study. J Epidemiol Community Health. 2007;61(1):40-7.
164. Matsumoto M, Ishikawa S, Kayaba K, Gotoh T, Nago N, Tsutsumi A, et al. Risk charts illustrating the 10-year risk of myocardial infarction among residents of Japanese rural communities: the JMS Cohort Study. J Epidemiol. 2009;19(2):94-100.
165. May M, Lawlor DA, Brindle P, Patel R, Ebrahim S. Cardiovascular disease risk assessment in older women: can we improve on Framingham? British Women's Heart and Health prospective cohort study. Heart. 2006;92(10):1396-401.

166. May M, Sterne JAC, Shipley M, Brunner E, d'Agostino R, Whincup P, et al. A coronary heart disease risk model for predicting the effect of potent antiretroviral therapy in HIV-1 infected men. Int J Epidemiol. 2007;36(6):1309-18.
167. McGeechan K, Liew G, Macaskill P, Irwig L, Klein R, Sharrett AR, et al. Risk prediction of coronary heart disease based on retinal vascular caliber (from the Atherosclerosis Risk In Communities [ARIC] Study). Am J Cardiol. 2008;102(1):58-63.
168. McGorrian C, Yusuf S, Islam S, Jung H, Rangarajan S, Avezum A, et al. Estimating modifiable coronary heart disease risk in multiple regions of the world: the INTERHEART Modifiable Risk Score. Eur Heart J. 2011;32(5):581-9.
169. McNeil JJ, Peeters A, Liew D, Lim S, Vos T. A model for predicting the future incidence of coronary heart disease within percentiles of coronary heart disease risk. J Cardiovasc Risk. 2001;8(1):31-7. Epub 2001/03/10.
170. Meigs JB, Nathan DM, D'Agostino Sr RB, Wilson PWF. Fasting and postchallenge glycemia and cardiovascular disease risk: The framingham offspring study. Diabetes Care. 2002;25(10):1845-50.
171. Melander O, Newton-Cheh C, Almgren P, Hedblad B, Berglund G, Engstrom G, et al. Novel and conventional biomarkers for prediction of incident cardiovascular events in the community. JAMA. 2009;302(1):49-57.
172. Menotti A, Farchi G, Seccareccia F. The prediction of coronary heart disease mortality as a function of major risk factors in over 30 000 men in the Italian RIFLE pooling Project. A comparison with the MRFIT primary screenees. The RIFLE research group. J Cardiovasc Risk. 1994;1(3):263-70. Epub 1994/10/01.
173. Menotti A, Keys A, Kromhout D, Nissinen A, Blackburn H, Fidanza F, et al. Twenty-five-year mortality from coronary heart disease and its prediction in five cohorts of middle-aged men in Finland, The Netherlands, and Italy. Prev Med. 1990;19(3):270-8. Epub 1990/05/01.
174. Menotti A, Lanti M, Agabiti-Rosei E, Carratelli L, Cavera G, Dormi A, et al. Riskard 2005. New tools for prediction of cardiovascular disease risk derived from Italian population studies. Nutr Metab Cardiovasc Dis. 2005;15(6):426-40.
175. Menotti A, Lanti M, Puddu PE, Carratelli L, Mancini M, Motolese M, et al. The risk functions incorporated in Riscard 2002: a software for the prediction of cardiovascular risk in the general population based on Italian data. Ital Heart J. 2002;3(2):114-21. Epub 2002/04/03.
176. Menotti A, Lanti M, Puddu PE, Mancini M, Zanchetti A, Cirillo M, et al. First risk functions for prediction of coronary and cardiovascular disease incidence in the Gubbio Population Study. Ital Heart J. 2000;1(6):394-9. Epub 2000/08/10.
177. Merry AHH, Boer JMA, Schouten LJ, Ambergen T, Steyerberg EW, Feskens EJM, et al. Risk prediction of incident coronary heart disease in The Netherlands: re-estimation and improvement of the SCORE risk function. Eur J Prev Cardiol. 2012;19(4):840-8.
178. Milne R, Gamble G, Whitlock G, Jackson R. Discriminative ability of a risk-prediction tool derived from the Framingham Heart Study compared with single risk factors. N Z Med J. 2003;116(1185):U663.

179. Milne R, Gamble G, Whitlock G, Jackson R. Framingham Heart Study risk equation predicts first cardiovascular event rates in New Zealanders at the population level. N Z Med J. 2003;116(1185):U662. Epub 2003/11/15.
180. Mitchell GF, Hwang S-J, Vasan RS, Larson MG, Pencina MJ, Hamburg NM, et al. Arterial stiffness and cardiovascular events: the Framingham Heart Study. Circulation. 2010;121(4):505-11.
181. Mohammadreza B, Farzad H, Davoud K, Fereidoun Prof AF. Prognostic significance of the complex "Visceral Adiposity Index" vs. simple anthropometric measures: Tehran lipid and glucose study. Cardiovasc Diabetol. 2012;11:20.
182. Mohlenkamp S, Lehmann N, Greenland P, Moebus S, Kalsch H, Schmermund A, et al. Coronary artery calcium score improves cardiovascular risk prediction in persons without indication for statin therapy. Atherosclerosis. 2011;215(1):229-36.
183. Mohlenkamp S, Lehmann N, Moebus S, Schmermund A, Dragano N, Stang A, et al. Quantification of coronary atherosclerosis and inflammation to predict coronary events and all-cause mortality. J Am Coll Cardiol. 2011;57(13):1455-64.
184. Moons KG, Bots ML, Salonen JT, Elwood PC, Freire de Concalves A, Nikitin Y, et al. Prediction of stroke in the general population in Europe (EUROSTROKE): Is there a role for fibrinogen and electrocardiography? J Epidemiol Community Health. 2002;56 Suppl 1:i30-6. Epub 2002/01/30.
185. Mora S, Redberg RF, Sharrett AR, Blumenthal RS. Enhanced risk assessment in asymptomatic individuals with exercise testing and Framingham risk scores. Circulation. 2005;112(11):1566-72.
186. Morrison AC, Bare LA, Chambless LE, Ellis SG, Malloy M, Kane JP, et al. Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. Am J Epidemiol. 2007;166(1):28-35.
187. Munir JA, Wu H, Bauer K, Bindeman J, Byrd C, O'Malley P, et al. Impact of coronary calcium on arterial age and coronary heart disease risk estimation using the MESA arterial age calculator. Atherosclerosis. 2010;211(2):467-70. Epub 2010/04/10.
188. Murphy TP, Dhangana R, Pencina MJ, D'Agostino RB, Sr. Ankle-brachial index and cardiovascular risk prediction: an analysis of 11,594 individuals with 10-year follow-up. Atherosclerosis. 2012;220(1):160-7.
189. Murphy TP, Dhangana R, Pencina MJ, Zafar AM, D'Agostino RB. Performance of current guidelines for coronary heart disease prevention: optimal use of the Framingham-based risk assessment. Atherosclerosis. 2011;216(2):452-7.
190. Nambi V, Boerwinkle E, Lawson K, Brautbar A, Chambless L, Franeschini N, et al. The 9p21 genetic variant is additive to carotid intima media thickness and plaque in improving coronary heart disease risk prediction in white participants of the Atherosclerosis Risk in Communities (ARIC) Study. Atherosclerosis. 2012;222(1):135-7.
191. Nambi V, Chambless L, Folsom AR, He M, Hu Y, Mosley T, et al. Carotid intima-media thickness and presence or absence of plaque improves prediction of coronary heart disease risk: the ARIC (Atherosclerosis Risk In Communities) study. J Am Coll Cardiol. 2010;55(15):1600-7.

192. Nambi V, Chambless L, He M, Folsom AR, Mosley T, Boerwinkle E, et al. Common carotid artery intima-media thickness is as good as carotid intima-media thickness of all carotid artery segments in improving prediction of coronary heart disease risk in the Atherosclerosis Risk in Communities (ARIC) study. Eur Heart J. 2012;33(2):183-90.
193. Nelson MR, Ramsay E, Ryan P, Willson K, Tonkin AM, Wing L, et al. A score for the prediction of cardiovascular events in the hypertensive aged. Am J Hypertens. 2012;25(2):190-4.
194. Nelson MR, Ryan P, Tonkin AM, Ramsay E, Willson K, Wing LWH, et al. Prediction of cardiovascular events in subjects in the second Australian National Blood Pressure study. Hypertension. 2010;56(1):44-8.
195. Nielsen M, Ganz M, Lauze F, Pettersen PC, de Bruijne M, Clarkson TB, et al. Distribution, size, shape, growth potential and extent of abdominal aortic calcified deposits predict mortality in postmenopausal women. BMC Cardiovasc Disord. 2010;10:56.
196. Nippon Data Research Group. Risk assessment chart for death from cardiovascular disease based on a 19-year follow-up study of a Japanese representative population. Circ J. 2006;70(10):1249-55.
197. Noda H, Maruyama K, Iso H, Dohi S, Terai T, Fujioka S, et al. Prediction of myocardial infarction using coronary risk scores among Japanese male workers: 3M Study. J Atheroscler Thromb. 2010;17(5):452-9.
198. Nordestgaard BG, Adourian AS, Freiberg JJ, Guo Y, Muntendam P, Falk E. Risk factors for near-term myocardial infarction in apparently healthy men and women. Clin Chem. 2010;56(4):559-67.
199. Novo S, Visconti CL, Amoroso GR, Corrado E, Fazio G, Muratori I, et al. Asymptomatic carotid lesions add to cardiovascular risk prediction. Eur J Cardiovasc Prev Rehabil. 2010;17(5):514-8.
200. Nozaki T, Sugiyama S, Koga H, Sugamura K, Ohba K, Matsuzawa Y, et al. Significance of a multiple biomarkers strategy including endothelial dysfunction to improve risk stratification for cardiovascular events in patients at high risk for coronary heart disease. J Am Coll Cardiol. 2009;54(7):601-8.
201. Odell PM, Anderson KM, Kannel WB. New models for predicting cardiovascular events. J Clin Epidemiol. 1994;47(6):583-92. Epub 1994/06/01.
202. Oksala N, Seppala I, Hernesniemi J, Lyytikainen L-P, Kahonen M, Makela K-M, et al. Complementary prediction of cardiovascular events by estimated apo- and lipoprotein concentrations in the working age population. The Health 2000 Study. Ann Med. 2013;45(2):141-8.
203. Olsen MH, Wachtell K, Ibsen H, Lindholm L, Kjeldsen SE, Omvik P, et al. Changes in subclinical organ damage vs. in Framingham risk score for assessing cardiovascular risk reduction during continued antihypertensive treatment: a LIFE substudy. J Hypertens. 2011;29(5):997-1004.
204. Onat A, Can G, Hergenc G, Ugur M, Yuksel H. Coronary disease risk prediction algorithm warranting incorporation of C-reactive protein in Turkish adults, manifesting sex difference. Nutr Metab Cardiovasc Dis. 2012;22(8):643-50.

205. Orford JL, Sesso HD, Stedman M, Gagnon D, Vokonas P, Gaziano JM. A comparison of the Framingham and European Society of Cardiology coronary heart disease risk prediction models in the normative aging study. Am Heart J. 2002;144(1):95-100. Epub 2002/07/03.
206. Panagiotakos DB, Fitzgerald AP, Pitsavos C, Pipilis A, Graham I, Stefanadis C. Statistical modelling of 10-year fatal cardiovascular disease risk in Greece: the HellenicSCORE (a calibration of the ESC SCORE project). Hellenic J Cardiol. 2007;48(2):55-63.
207. Panagiotakos DB, Pitsavos C, Stefanadis C. Inclusion of dietary evaluation in cardiovascular disease risk prediction models increases accuracy and reduces bias of the estimations. Risk Anal. 2009;29(2):176-86.
208. Pandya A, Weinstein MC, Gaziano TA. A comparative assessment of non-laboratory-based versus commonly used laboratory-based cardiovascular disease risk scores in the NHANES III population. PLoS ONE. 2011;6(5):e20416.
209. Park Y, Lim J, Lee J, Kim SG. Erythrocyte fatty acid profiles can predict acute non-fatal myocardial infarction. Br J Nutr. 2009;102(9):1355-61. Epub 2009/06/10.
210. Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, Ridker PM. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. Ann Intern Med. 2009;150(2):65-72.
211. Paynter NP, Chasman DI, Pare G, Buring JE, Cook NR, Miletich JP, et al. Association between a literature-based genetic risk score and cardiovascular events in women. JAMA. 2010;303(7):631-7. Epub 2010/02/18.
212. Paynter NP, Mazer NA, Pradhan AD, Gaziano JM, Ridker PM, Cook NR. Cardiovascular risk prediction in diabetic men and women using hemoglobin A1c vs diabetes as a high-risk equivalent. Arch Intern Med. 2011;171(19):1712-8.
213. Pencina MJ, D'Agostino RB, Sr., Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the framingham heart study. Circulation. 2009;119(24):3078-84.
214. Petersson U, Ostgren CJ, Brudin L, Nilsson PM. A consultation-based method is equal to SCORE and an extensive laboratory-based method in predicting risk of future cardiovascular disease. Eur J Cardiovasc Prev Rehabil. 2009;16(5):536-40.
215. Plichart M, Celermajer DS, Zureik M, Helmer C, Jouven X, Ritchie K, et al. Carotid intima-media thickness in plaque-free site, carotid plaques and coronary heart disease risk prediction in older adults. The Three-City Study. Atherosclerosis. 2011;219(2):917-24.
216. Pocock SJ, McCormack V, Gueyffier F, Boutitie F, Fagard RH, Boissel JP. A score for predicting risk of death from cardiovascular disease in adults with raised blood pressure, based on individual patient data from randomised controlled trials. BMJ. 2001;323(7304):75-81. Epub 2001/07/14.
217. Poels MMF, Steyerberg EW, Wieberdink RG, Hofman A, Koudstaal PJ, Ikram MA, et al. Assessment of cerebral small vessel disease predicts individual stroke risk. J Neurol Neurosurg Psychiatry. 2012;83(12):1174-9.

218. Polak JF, Pencina MJ, Pencina KM, O'Donnell CJ, Wolf PA, D'Agostino RB, Sr. Carotid-wall intima-media thickness and cardiovascular events. N Engl J Med. 2011;365(3):213-21.
219. Polonsky TS, McClelland RL, Jorgensen NW, Bild DE, Burke GL, Guerci AD, et al. Coronary artery calcium score and risk classification for coronary heart disease prediction. JAMA. 2010;303(16):1610-6.
220. Prati P, Tosetto A, Casaroli M, Bignamini A, Canciani L, Bornstein N, et al. Carotid plaque morphology improves stroke risk prediction: usefulness of a new ultrasonographic score. Cerebrovasc Dis. 2011;31(3):300-4.
221. Prugger C, Luc G, Haas B, Arveiler D, Machez E, Ferrieres J, et al. Adipocytokines and the risk of ischemic stroke: the PRIME Study. Ann Neurol. 2012;71(4):478-86.
222. Qiao Q, Gao W, Laatikainen T, Vartiainen E. Layperson-oriented vs. clinical-based models for prediction of incidence of ischemic stroke: National FINRISK Study. Int J Stroke. 2012;7(8):662-8.
223. Rachas A, Raffaitin C, Barberger-Gateau P, Helmer C, Ritchie K, Tzourio C, et al. Clinical usefulness of the metabolic syndrome for the risk of coronary heart disease does not exceed the sum of its individual components in older men and women. The Three-City (3C) Study. Heart. 2012;98(8):650-5.
224. Ramachandran S, French JM, Vanderpump MP, Croft P, Neary RH. Using the Framingham model to predict heart disease in the United Kingdom: retrospective study. BMJ. 2000;320(7236):676-7. Epub 2000/03/11.
225. Ramsay SE, Morris RW, Whincup PH, Papacosta AO, Thomas MC, Wannamethee SG. Prediction of coronary heart disease risk by Framingham and SCORE risk assessments varies by socioeconomic position: results from a study in British men. Eur J Cardiovasc Prev Rehabil. 2011;18(2):186-93.
226. Rana JS, Cote M, Despres JP, Sandhu MS, Talmud PJ, Ninio E, et al. Inflammatory biomarkers and the prediction of coronary events among people at intermediate risk: the EPIC-Norfolk prospective population study. Heart. 2009;95(20):1682-7.
227. Reissigova J, Zvarova J. The Framingham risk function underestimated absolute coronary heart disease risk in Czech men. Methods Inf Med. 2007;46(1):43-9.
228. Riddell T, Wells S, Jackson R, Lee A-W, Crengle S, Bramley D, et al. Performance of Framingham cardiovascular risk scores by ethnic groups in New Zealand: PREDICT CVD-10. N Z Med J. 2010;123(1309):50-61.
229. Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. JAMA. 2007;297(6):611-9.
230. Ridker PM, Paynter NP, Rifai N, Gaziano JM, Cook NR. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. Circulation. 2008;118(22):2243-51, 4p following 51.
231. Rifkin DE, Ix JH, Wassel CL, Criqui MH, Allison MA. Renal artery calcification and mortality among clinically asymptomatic adults. J Am Coll Cardiol. 2012;60(12):1079-85.
232. Rodondi N, Locatelli I, Aujesky D, Butler J, Vittinghoff E, Simonsick E, et al. Framingham risk score and alternatives for prediction of coronary heart disease in older adults. PLoS ONE. 2012;7(3):e34287.

233. Root M, Smith T. Prescribe by risk: the utility of a biomarker-based risk calculation in disease management to prevent heart disease. Dis Manag. 2005;8(2):106-13.
234. Rutten JHW, Mattace-Raso FUS, Steyerberg EW, Lindemans J, Hofman A, Wieberdink RG, et al. Amino-terminal pro-B-type natriuretic peptide improves cardiovascular and cerebrovascular risk prediction in the population: the Rotterdam study. Hypertension. 2010;55(3):785-91.
235. Ruwald MH, Ruwald AC, Jons C, Lamberts M, Hansen ML, Vinther M, et al. Evaluation of the chads2 risk score on short- and long-term all-cause and cardiovascular mortality after syncope. Clin Cardiol. 2013;36(5):262-8.
236. Sacco RL, Khatri M, Rundek T, Xu Q, Gardener H, Boden-Albala B, et al. Improving global vascular risk prediction with behavioral and anthropometric factors. The multiethnic NOMAS (Northern Manhattan Cohort Study). J Am Coll Cardiol. 2009;54(24):2303-11.
237. Saidj M, Jorgensen T, Prescott E, Borglykke A. Poor predictive ability of the risk chart SCORE in a Danish population. Dan Med J. 2013;60(5).
238. Saunders JT, Nambi V, de Lemos JA, Chambless LE, Virani SS, Boerwinkle E, et al. Cardiac troponin T measured by a highly sensitive assay predicts coronary heart disease, heart failure, and mortality in the Atherosclerosis Risk in Communities Study. Circulation. 2011;123(13):1367-76.
239. Scheltens T, Verschuren WMM, Boshuizen HC, Hoes AW, Zuithoff NP, Bots ML, et al. Estimation of cardiovascular risk: a comparison between the Framingham and the SCORE model in people under 60 years of age. Eur J Cardiovasc Prev Rehabil. 2008;15(5):562-6.
240. Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB, Sr., et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. Lancet. 2009;373(9665):739-45. Epub 2009/03/03.
241. Schottker B, Muller H, Rothenbacher D, Brenner H. Fasting plasma glucose and HbA1c in cardiovascular risk prediction: A sex-specific comparison in individuals without diabetes mellitus. Diabetologia. 2013;56(1):92-100.
242. Sehestedt T, Jeppesen J, Hansen TW, Wachtell K, Ibsen H, Torp-Petersen C, et al. Risk prediction is improved by adding markers of subclinical organ damage to SCORE. Eur Heart J. 2010;31(7):883-91.
243. Sever PS, Poulter NR, Chang CL, Hingorani A, Thom SA, Hughes AD, et al. Evaluation of C-reactive protein prior to and on-treatment as a predictor of benefit from atorvastatin: observations from the Anglo-Scandinavian Cardiac Outcomes Trial. Eur Heart J. 2012;33(4):486-94. Epub 2011/07/30.
244. Shah S, Casas JP, Gaunt TR, Cooper J, Drenos F, Zabaneh D, et al. Influence of common genetic variation on blood lipid levels, cardiovascular risk, and coronary events in two British prospective cohort studies. Eur Heart J. 2013;34(13):972-81.
245. Shaper AG, Pocock SJ, Phillips AN, Walker M. Identifying men at high risk of heart attacks: strategy for use in general practice. Br Med J (Clin Res Ed). 1986;293(6545):474-9. Epub 1986/08/23.

246. Shara NM, Wang H, Valaitis E, Pehlivanova M, Carter EA, Resnick HE, et al. Comparison of estimated glomerular filtration rates and albuminuria in predicting risk of coronary heart disease in a population with high prevalence of diabetes mellitus and renal disease. Am J Cardiol. 2011;107(3):399-405.
247. Simmons RK, Coleman RL, Price HC, Holman RR, Khaw K-T, Wareham NJ, et al. Performance of the UK Prospective Diabetes Study Risk Engine and the Framingham Risk Equations in Estimating Cardiovascular Disease in the EPIC- Norfolk Cohort. Diabetes Care. 2009;32(4):708-13.
248. Simmons RK, Sharp S, Boekholdt SM, Sargeant LA, Khaw K-T, Wareham NJ, 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? Arch Intern Med. 2008;168(11):1209-16.
249. Simons LA, Simons J, Friedlander Y, McCallum J, Palaniappan L. Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study. Med J Aust. 2003;178(3):113-6. Epub 2003/02/01.
250. Sivapalaratnam S, Boekholdt SM, Trip MD, Sandhu MS, Luben R, Kastelein JJP, et al. Family history of premature coronary heart disease and risk prediction in the EPIC-Norfolk prospective population study. Heart. 2010;96(24):1985-9.
251. Smink PA, Lambers Heerspink HJ, Gansevoort RT, de Jong PE, Hillege HL, Bakker SJL, et al. Albuminuria, estimated GFR, traditional risk factors, and incident cardiovascular disease: the PREVEND (Prevention of Renal and Vascular Endstage Disease) study. Am J Kidney Dis. 2012;60(5):804-11.
252. Smith JG, Newton-Cheh C, Almgren P, Struck J, Morgenthaler NG, Bergmann A, et al. Assessment of conventional cardiovascular risk factors and multiple biomarkers for the prediction of incident heart failure and atrial fibrillation. J Am Coll Cardiol. 2010;56(21):1712-9.
253. Stein PK, Barzilay JI. Relationship of abnormal heart rate turbulence and elevated CRP to cardiac mortality in low, intermediate, and high-risk older adults. J Cardiovasc Electrophysiol. 2011;22(2):122-7.
254. Stenlund H, Lonnberg G, Jenkins P, Norberg M, Persson M, Messner T, et al. Fewer deaths from cardiovascular disease than expected from the Systematic Coronary Risk Evaluation chart in a Swedish population. Eur J Cardiovasc Prev Rehabil. 2009;16(3):321-4.
255. Stern MP, Williams K, Gonzalez-Villalpando C, Hunt KJ, Haffner SM. Does the metabolic syndrome improve identification of individuals at risk of type 2 diabetes and/or cardiovascular disease? Diabetes Care. 2004;27(11):2676-81.
256. Stork S, Feelders RA, van den Beld AW, Steyerberg EW, Savelkoul HFJ, Lamberts SWJ, et al. Prediction of mortality risk in the elderly. Am J Med. 2006;119(6):519-25.
257. Suka M, Sugimori H, Yoshida K. Application of the updated Framingham risk score to Japanese men. Hypertens Res. 2001;24(6):685-9. Epub 2002/01/05.
258. Talmud PJ, Cooper JA, Palmen J, Lovering R, Drenos F, Hingorani AD, et al. Chromosome 9p21.3 coronary heart disease locus genotype and prospective risk of CHD in healthy middle-aged men. Clin Chem. 2008;54(3):467-74.

259. Tanabe N, Iso H, Okada K, Nakamura Y, Harada A, Ohashi Y, et al. Serum total and non-high-density lipoprotein cholesterol and the risk prediction of cardiovascular events - the JALS-ECC. Circ J. 2010;74(7):1346-56. Epub 2010/06/08.
260. Teramoto T, Ohashi Y, Nakaya N, Yokoyama S, Mizuno K, Nakamura H, et al. Practical risk prediction tools for coronary heart disease in mild to moderate hypercholesterolemia in Japan: originated from the MEGA study data. Circ J. 2008;72(10):1569-75.
261. Thanassoulis G, Peloso GM, Pencina MJ, Hoffmann U, Fox CS, Cupples LA, et al. A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium the framingham heart study. Circ Cardiovasc Genet. 2012;5(1):113-21.
262. Thomsen TF, Davidsen M, Ibsen H, Jorgensen T, Jensen G, Borch-Johnsen K. A new method for CHD prediction and prevention based on regional risk scores and randomized clinical trials; PRECARD and the Copenhagen Risk Score. J Cardiovasc Risk. 2001;8(5):291-7. Epub 2001/11/10.
263. Thorsen RD, Jacobs DR, Jr., Grimm RH, Jr., Keys A, Taylor H, Blackburn H. Preventive cardiology in practice: a device for risk estimation and counseling in coronary disease. Prev Med. 1979;8(5):548-56. Epub 1979/09/01.
264. Tohidi M, Hadaegh F, Harati H, Azizi F. C-reactive protein in risk prediction of cardiovascular outcomes: Tehran Lipid and Glucose Study. Int J Cardiol. 2009;132(3):369-74.
265. Truelsen T, Lindenstrom E, Boysen G. Comparison of probability of stroke between the Copenhagen City Heart Study and the Framingham Study. Stroke. 1994;25(4):802-7. Epub 1994/04/01.
266. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. J Chronic Dis. 1967;20(7):511-24. Epub 1967/07/01.
267. Tsang TS, Barnes ME, Gersh BJ, Takemoto Y, Rosales AG, Bailey KR, et al. Prediction of risk for first age-related cardiovascular events in an elderly population: the incremental value of echocardiography. J Am Coll Cardiol. 2003;42(7):1199-205. Epub 2003/10/03.
268. Tsimikas S, Mallat Z, Talmud PJ, Kastelein JJP, Wareham NJ, Sandhu MS, et al. Oxidation-specific biomarkers, lipoprotein(a), and risk of fatal and nonfatal coronary events. J Am Coll Cardiol. 2010;56(12):946-55.
269. Tsimikas S, Willeit P, Willeit J, Santer P, Mayr M, Xu Q, et al. Oxidation-specific biomarkers, prospective 15-year cardiovascular and stroke outcomes, and net reclassification of cardiovascular events. J Am Coll Cardiol. 2012;60(21):2218-29.
270. Tunstall-Pedoe H. The Dundee coronary risk-disk for management of change in risk factors. BMJ. 1991;303(6805):744-7. Epub 1991/09/28.
271. Tunstall-Pedoe H, Woodward M, estimation Sgor. By neglecting deprivation, cardiovascular risk scoring will exacerbate social gradients in disease. Heart. 2006;92(3):307-10.
272. Ulmer H, Kollerits B, Kelleher C, Diem G, Concin H. Predictive accuracy of the SCORE risk function for cardiovascular disease in clinical practice: a prospective evaluation of 44 649 Austrian men and women. Eur J Cardiovasc Prev Rehabil. 2005;12(5):433-41.

273. Vaidya D, Yanek LR, Moy TF, Pearson TA, Becker LC, Becker DM. Incidence of coronary artery disease in siblings of patients with premature coronary artery disease: 10 years of follow-up. Am J Cardiol. 2007;100(9):1410-5.
274. van der Heijden AAWA, Ortegon MM, Niessen LW, Nijpels G, Dekker JM. Prediction of coronary heart disease risk in a general, pre-diabetic, and diabetic population during 10 years of follow-up: accuracy of the Framingham, SCORE, and UKPDS risk functions: The Hoorn Study. Diabetes Care. 2009;32(11):2094-8.
275. van Dis I, Kromhout D, Geleijnse JM, Boer JMA, Verschuren WMM. Evaluation of cardiovascular risk predicted by different SCORE equations: the Netherlands as an example. Eur J Cardiovasc Prev Rehabil. 2010;17(2):244-9.
276. Veeranna V, Zalawadiya SK, Niraj A, Pradhan J, Ference B, Burack RC, et al. Homocysteine and reclassification of cardiovascular disease risk. J Am Coll Cardiol. 2011;58(10):1025-33.
277. Venskutonyte L, Ryden L, Nilsson G, Ohrvik J. Mortality prediction in the elderly by an easily measured metabolic index. Diab Vasc Dis Res. 2012;9(3):226-33. Epub 2012/01/27.
278. Vergnaud AC, Bertrais S, Galan P, Hercberg S, Czernichow S. Ten-year risk prediction in French men using the Framingham coronary score: results from the national SU.VI.MAX cohort. Prev Med. 2008;47(1):61-5.
279. Verwoert GC, Elias-Smale SE, Rizopoulos D, Koller MT, Steyerberg EW, Hofman A, et al. Does aortic stiffness improve the prediction of coronary heart disease in elderly? The Rotterdam Study. J Hum Hypertens. 2012;26(1):28-34.
280. Villines TC, Taylor AJ. Multi-ethnic study of atherosclerosis arterial age versus framingham 10-year or lifetime cardiovascular risk. Am J Cardiol. 2012;110(11):1627-30.
281. Vlismas K, Panagiotakos DB, Pitsavos C, Chrysohoou C, Skoumas Y, Stavrinos V, et al. The role of dietary and socioeconomic status assessment on the predictive ability of the HellenicSCORE. Hellenic J Cardiol. 2011;52(5):391-8.
282. Voko Z, Hollander M, Koudstaal PJ, Hofman A, Breteler MMB. How do American stroke risk functions perform in a Western European population? Neuroepidemiology. 2004;23(5):247-53.
283. Voss R, Cullen P, Schulte H, Assmann G. Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Munster Study (PROCAM) using neural networks. Int J Epidemiol. 2002;31(6):1253-62; discussion 62-64. Epub 2003/01/24.
284. Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. N Engl J Med. 2006;355(25):2631-9.
285. Wang Z, Hoy WE. Is the Framingham coronary heart disease absolute risk function applicable to Aboriginal people? Med J Aust. 2005;182(2):66-9. Epub 2005/01/18.
286. Wannamethee SG, Shaper AG, Lennon L, Morris RW. Metabolic syndrome vs Framingham Risk Score for prediction of coronary heart disease, stroke, and type 2 diabetes mellitus. Arch Intern Med. 2005;165(22):2644-50.

287. Weiner DE, Tighiouart H, Griffith JL, Elsayed E, Levey AS, Salem DN, et al. Kidney disease, Framingham risk scores, and cardiac and mortality outcomes. Am J Med. 2007;120(6):552.e1-8.
288. Wilson PW, Castelli WP, Kannel WB. Coronary risk prediction in adults (the Framingham Heart Study). Am J Cardiol. 1987;59(14):91G-4G. Epub 1987/05/29.
289. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837-47. Epub 1998/05/29.
290. Wilson PWF, Nam B-H, Pencina M, D'Agostino RB, Sr., Benjamin EJ, O'Donnell CJ. C-reactive protein and risk of cardiovascular disease in men and women from the Framingham Heart Study. Arch Intern Med. 2005;165(21):2473-8.
291. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham Study. Stroke. 1991;22(3):312-8. Epub 1991/03/01.
292. Woodward M, Brindle P, Tunstall-Pedoe H, estimation Sgor. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). Heart. 2007;93(2):172-6.
293. Woodward M, Tunstall-Pedoe H, Batty GD, Tavendale R, Hu FB, Czernichow S. The prognostic value of adipose tissue fatty acids for incident cardiovascular disease: results from 3944 subjects in the Scottish Heart Health Extended Cohort Study. Eur Heart J. 2011;32(11):1416-23.
294. Woodward M, Tunstall-Pedoe H, Rumley A, Lowe GDO. Does fibrinogen add to prediction of cardiovascular disease? Results from the Scottish Heart Health Extended Cohort Study. Br J Haematol. 2009;146(4):442-6.
295. Woodward M, Welsh P, Rumley A, Tunstall-Pedoe H, Lowe GDO. Do inflammatory biomarkers add to the discrimination of cardiovascular disease after allowing for social deprivation? Results from a 10-year cohort study in Glasgow, Scotland. Eur Heart J. 2010;31(21):2669-75.
296. Wormser D, Kaptoge S, Di Angelantonio E, Wood AM, Pennells L, Thompson A, et al. Separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease: Collaborative analysis of 58 prospective studies. The Lancet. 2011;377(9771):1085-95.
297. Wu Y, Liu X, Li X, Li Y, Zhao L, Chen Z, et al. Estimation of 10-year risk of fatal and nonfatal ischemic cardiovascular diseases in Chinese adults. Circulation. 2006;114(21):2217-25.
298. Wu Y, Zhang L, Yuan X, Wu Y, Yi D. Quantifying links between stroke and risk factors: a study on individual health risk appraisal of stroke in a community of Chongqing. Neurol Sci. 2011;32(2):211-9.
299. Xie W, Liang L, Zhao L, Shi P, Yang Y, Xie G, et al. Combination of carotid intima-media thickness and plaque for better predicting risk of ischaemic cardiovascular events. Heart. 2011;97(16):1326-31.
300. Yip YB, Wong TKS, Chung JWY, Ko SKK, Sit JWH, Chan TMF. Cardiovascular disease: application of a composite risk index from the Telehealth System in a district community. Public Health Nurs. 2004;21(6):524-32.

301. Zhang X-F, Attia J, D'Este C, Yu X-H, Wu X-G. A risk score predicted coronary heart disease and stroke in a Chinese cohort. J Clin Epidemiol. 2005;58(9):951-8.
302. Zomer E, Owen A, Magliano DJ, Liew D, Reid C. Validation of two Framingham cardiovascular risk prediction algorithms in an Australian population: the 'old' versus the 'new' Framingham equation. Eur J Cardiovasc Prev Rehabil. 2011;18(1):115-20.

Chapter 7

Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement

Pauline Heus
Johanna AAG Damen
Romin Pajouheshnia
Rob JPM Scholten
Johannes B Reitsma
Gary S Collins
Douglas G Altman
Karel GM Moons
Lotty Hooft

Submitted

Abstract

Background: As complete reporting is essential to judge the validity and applicability of multivariable prediction models, a guideline for the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) was introduced. We assessed the completeness of reporting of prediction model studies published just before the introduction of the TRIPOD statement, to refine and tailor the implementation strategy.

Methods: Within each of 37 clinical domains, 10 journals with the highest journal impact factor were selected. A PubMed search was performed to identify prediction model studies published before the launch of TRIPOD (May 2014) in these journals. Eligible publications reported on the development or external validation of a multivariable prediction model (either diagnostic or prognostic), or on the incremental value of adding a predictor to an existing model.

Results: We included 146 publications (84% prognostic), from which we assessed 170 models: 73 (43%) model development, 43 (25%) external validation, 33 (19%) incremental value, and 21 (12%) combined development and external validation of the same model. Overall, publications adhered to a median of 44% (25^{th}–75^{th} percentile: 35% to 52%) of TRIPOD items, with 44% (36% to 53%) for prognostic and 41% (34% to 48%) for diagnostic models. TRIPOD items that were completely reported for less than 25% of the models concerned title (5%), blinding of predictor assessment (6%), abstract (8%), comparison of development and validation data (11%), model updating (14%), model performance (15%), model specification (17%), characteristics of participants (21%), model performance measures (methods) (22%), and model building procedures (24%). Most often reported were TRIPOD items regarding overall interpretation (96%), source of data (95%), and risk groups (90%).

Conclusions: More than half of the items considered essential for transparent reporting were not fully addressed in publications of multivariable prediction model studies. Essential information for using a model in individual risk prediction, i.e. model specifications and model performance, was incomplete for over 80% of the models. Items that require improvement are title, abstract, and model building procedures, as they are crucial for identification and external validation of prediction models.

Introduction

Multivariable prediction models (risk scores or prediction rules) estimate an individual's probability or risk that a specific disease or condition is present (diagnostic models) or that a specific event will occur in the future (prognostic models) based on multiple characteristics or pieces of information of that individual.[1] Such models are increasingly used by healthcare providers to support clinical decision making or to inform patients or relatives. Studies about prediction models may address the development of a new model, validation of an existing, previously developed model in other individuals (with or without adjusting or updating the model to the validation setting), or a combination of these two.[2-5] Some prediction model studies evaluate the addition of a single predictor to an existing model (incremental value).[4]

In addition to appropriate design, conduct and analysis, reporting of prediction model studies should be complete and accurate. Complete reporting of research facilitates study replication, assessment of the study validity (risk of bias), interpretation of the results, and judgement of applicability of the study results (e.g. the prediction model itself) to other individuals or settings. Clinicians and other stakeholders can only use previously developed and validated prediction models when all relevant information is available for calculating predicted risks at an individual level. High quality information about prediction model studies is therefore essential.

Previous systematic reviews showed that within different clinical domains the quality of reporting of prediction models is suboptimal.[6-11] To improve the reporting of studies of prediction models, a guideline for the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) was launched in January 2015 in over 10 medical journals.[12,13] The TRIPOD statement is a checklist of 22 items considered essential for informative reporting of prediction model studies. Both diagnostic and prognostic prediction model studies are covered by the TRIPOD statement, and the checklist can be used for all types of prediction model studies (development, external validation, and incremental value) within all clinical domains.

In this comprehensive literature review, we assessed the completeness of reporting of prediction model studies that were published just before the introduction of the TRIPOD statement. Our results provide key clues to further refine and tailor the implementation strategy of the TRIPOD statement.

Methods

Identification of prediction model studies

To cover a wide range of clinical domains we started with 37 subject categories (2012 Journal Citation Reports®)[14] from which we selected the 10 journals with the highest Journal Impact Factor (Supplemental Table 1). After deduplication, 341 unique journals remained. We performed a search in PubMed to identify prediction model studies published in these journals before the launch of TRIPOD (May 2014), using a validated search filter for identifying prognostic and diagnostic prediction studies (Supplemental Table 2).[15]

Eligible publications described the development or external validation of a multivariable prediction model (either diagnostic or prognostic), or evaluated the incremental value of adding a predictor to an existing model.[1-5,16] We excluded so-called prognostic factor or predictor finding studies, as well as studies evaluating the impact of the use of a prediction model on management or patient outcomes.[3,7,17] We excluded prediction model studies using non-regression techniques (e.g. classification trees, neural networks and machine learning) or pharmacokinetic models. Titles and abstracts of the retrieved publications were screened by one of two authors (JAAGD or PH). After reading the full text report, they judged whether to include or exclude a potentially eligible publication. Any doubts regarding definitive eligibility were discussed, if necessary, with a third author. If we were not able to retrieve the full text of a publication via our institutions, it was excluded.

Data-extraction

For each included publication we recorded the journal impact factor (2012 Journal Citation Reports®)[14] clinical domain, and whether the purpose of prediction was diagnostic or prognostic. Furthermore, we classified publications into four types of prediction model studies: development, external validation, incremental value, or combination of development and external validation of the same model. A publication could be categorized as more than one type of prediction model study. For example, if a publication reported on both development and external validation, but of different models, it was classified as development as well as external validation. If a publication included multiple prediction model studies of the same type, e.g. two models were developed, we extracted data for only one model. If there was no primary model, we used the model that was studied in the largest sample. Information about study design, sample size, number of predictors in the final model, and predicted outcome was extracted for all included prediction models.

To judge the completeness of the reporting, we transformed items of the TRIPOD statement (Box 1) into a data-extraction form, which was piloted extensively to ensure consistent extraction of the data. The TRIPOD statement consists of 22 main items, of

which ten are divided in two (items 3, 4, 6, 7, 14, 15, and 19), three (items 5 and 13), or five (item 10) sub items.(12, 13) For TRIPOD items (main or sub items, hereafter just called items) containing multiple reporting elements we extracted information regarding each of these elements. For example, for item 4b *"Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up."* we used three data extraction elements to record information regarding 1) the start of accrual, 2) end of accrual, and 3) end of follow-up.

For each data extraction element we judged whether the requested information was available in the publication. If a publication reported both the development and external validation of the same prediction model, we extracted data on the reporting of either separately, and subsequently combined the extracted information for each data extraction element.

Three authors extracted data (JAAGD, PH, RP). If the authors disagreed or were unsure about the reporting of a data extraction element, it was discussed in consensus meetings with the other co-authors.

Analyses

Based on the extracted data elements, we first determined whether the reporting of each TRIPOD item was complete (definition see below). We then calculated overall scores for completeness of reporting per model, per publication, and per item of the TRIPOD statement (across models).

Completeness of reporting of each TRIPOD item

The reporting of a TRIPOD item was judged to be complete if the requested information for all elements of that particular TRIPOD item was present. For elements belonging to TRIPOD items 4b, 5a, 5c, 6a, 7a and 10a we considered a reference to information in another article acceptable. If an element was not applicable to a specific model, for example follow-up might be not relevant in a diagnostic prediction model study (item 4b), or blinding was a non-issue (e.g. if the predicted outcome was for example overall mortality) (items 6b and 7b), this element was regarded as being reported.

Box 1: Items of the TRIPOD statement

Title and abstract

1. **Title (D; V)**: identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.

2. **Abstract (D; V)**: provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.

Introduction

3. **Background and objectives**:

a. **(D; V)** Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.

b. **(D; V)** Specify the objectives, including whether the study describes the development or validation of the model or both.

Methods

4. **Source of data**:

a. **(D; V)** Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.

b. **(D; V)** Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.

5. **Participants**:

a. **(D; V)** Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.

b. **(D; V)** Describe eligibility criteria for participants.

c. **(D; V)** Give details of treatments received, if relevant.

6. **Outcome**:

a. **(D; V)** Clearly define the outcome that is predicted by the prediction model, including how and when assessed.

b. **(D; V)** Report any actions to blind assessment of the outcome to be predicted.

7. **Predictors**:

a. **(D; V)** Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.

b. **(D; V)** Report any actions to blind assessment of predictors for the outcome and other predictors.

8. **Sample size (D; V)**: explain how the study size was arrived at.

9. **Missing data (D; V)**: Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.

Box 1: Items of the TRIPOD statement

10.	**Statistical analysis methods**:
a.	**(D)** Describe how predictors were handled in the analyses.
b.	**(D)** Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.
c.	**(V)**For validation, describe how the predictions were calculated.
d.	**(D; V)** Specify all measures used to assess model performance and, if relevant, to compare multiple models.
e.	**(V)** Describe any model updating (e.g., recalibration) arising from the validation, if done.
11.	**Risk groups (D; V)**: Provide details on how risk groups were created, if done.
12.	**Development vs. validation (V)**: for validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.
Results	
13.	**Participants**:
a.	**(D; V)** Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.
b.	**(D; V)** Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.
c.	**(V)** For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).
14.	**Model development**:
a.	**(D)** Specify the number of participants and outcome events in each analysis.
b.	**(D)**If done, report the unadjusted association between each candidate predictor and outcome.
15.	**Model specification**:
a.	**(D)** Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).
b.	**(D)** Explain how to the use the prediction model.
16.	**Model performance (D;V)**: report performance measures (with CIs) for the prediction model.
17.	**Model-updating (V)**: if done, report the results from any model updating (i.e., model specification, model performance).
Discussion	
18.	**Limitations (D;V)**: discuss any limitations of the study (such as non-representative sample, few events per predictor, missing data).

Box 1: Items of the TRIPOD statement

19.	**Interpretation**:
a.	**(V)** For validation, discuss the results with reference to performance in the development data, and any other validation data.
b.	**(D;V)**Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.
20.	**Implications (D;V)**: discuss the potential clinical use of the model and implications for future research.
Other information	
21.	**Supplementary information (D;V)**: provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.
22.	**Funding (D;V)**: give the source of funding and the role of the funders for the present study.

D;V: item relevant to both development and external validation; *D*: item only relevant to development; *V*: item only relevant to external validation

Overall completeness of reporting per model

To calculate overall completeness of reporting for each included model we divided the number of completely reported TRIPOD items by the total number of TRIPOD items for that model. The total number of TRIPOD items varies per type of prediction model study, as six of the TRIPOD items only apply to development of a prediction model (10a, 10b, 14a, 14b, 15a, and 15b) and six only to external validation (10c, 10e, 12, 13c, 17, and 19a). This resulted in a total number of 31 TRIPOD items for the reporting of either development or external validation of a prediction model, 37 for the combined reporting of development and external validation of the same prediction model, and 36 for reporting incremental value.

Five items of the TRIPOD statement include an 'if done' or 'if applicable' statement (items 5c, 10e, 11, 14b and 17). If we considered such an item not applicable for a particular study, it was excluded when calculating the completeness of reporting (both in numerator and denominator). Furthermore, item 21 of the TRIPOD statement was excluded from all calculations, as it refers to whether supplementary material was provided.

Overall completeness of reporting per publication

The overall reporting per publication equals the reporting per model (see previous paragraph) for publications classified as either development, external validation, incremental value, or combined development and external validation of the same model. For publications classified as more than one type of prediction model study, for example development of a model and external validation of a different model, we combined the reporting of the different prediction model types within that publication. Reporting was considered complete when the reporting of the different types of prediction model

studies was complete, except for TRIPOD items 3a and 18-20, for which complete reporting for either type was considered sufficient.

We used linear regression to investigate possible relationships between completeness of reporting per publication as dependent variable, and sample size, journal impact factor, number of predictors in the final model, and prospective study design (as dichotomous variable, yes/no) as independent variables.

Overall completeness of reporting per item of the TRIPOD statement

We assessed the overall completeness of reporting of individual items of the TRIPOD statement by dividing the number of models with complete reporting of a particular TRIPOD item by the total number of models in which that item was applicable.

Results

We included a total of 146 publications (Figure 1). Most publications (122 [84%]) reported prognostic models. From the 146 publications we scored the reporting of 170 prediction models: 73 (43%) concerned model development, 43 (25%) external validation of an existing model, 33 (19%) incremental value of adding a predictor to a model, and 21 (12%) a combination of development and external validation of the same model.

The three clinical domains with most publications of prediction models were critical care medicine (18 [11%]), obstetrics and gynaecology (15 [9%]), and gastroenterology and hepatology (12 [7%]). The median journal impact factor of the publications was 5.3 (25^{th}-75^{th} percentile [P_{25}-P_{75}]: 4.0-7.1). Median sample size of the populations in which a model was studied was 450 (P_{25}-P_{75}: 200-2005). In the final models a median of 5 (P_{25}-P_{75}: 3-8) predictors were included and in 23 models (16%) all-cause mortality was the predicted outcome.

Completeness of reporting per publication

Overall, publications adhered to between 16% to 82% of the items of the TRIPOD statement with a median of 44% (P_{25}-P_{75}: 35%-52%) (Figure 2). The reporting quality for prognostic and diagnostic prediction models was comparable, with median adherence of 44% (P_{25}-P_{75}: 36%-53%) and 41% (P_{25}-P_{75}: 34%-48%), respectively. The most complete reporting was seen for the combined reporting of development and external validation of the same model (49%, P_{25}-P_{75}: 35%-54%), followed by the reporting of model development (43%; P_{25}-P_{75}: 35%-53%), external validation (43%; P_{25}-P_{75}: 37%-54%), and incremental value (40%; P_{25}-P_{75}: 33%-49%). No associations were found between completeness of reporting and sample size, journal impact factor, number of predictors in the final model, and prospective study design (data not shown).

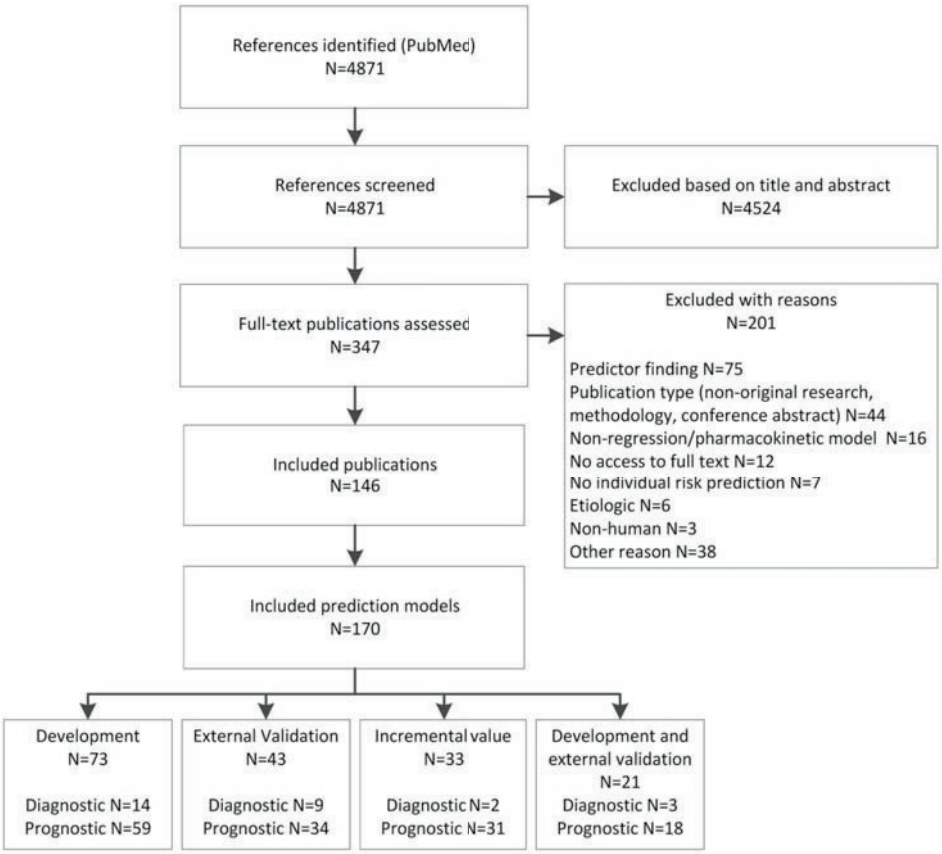


Figure 1: Flow diagram of selection procedure

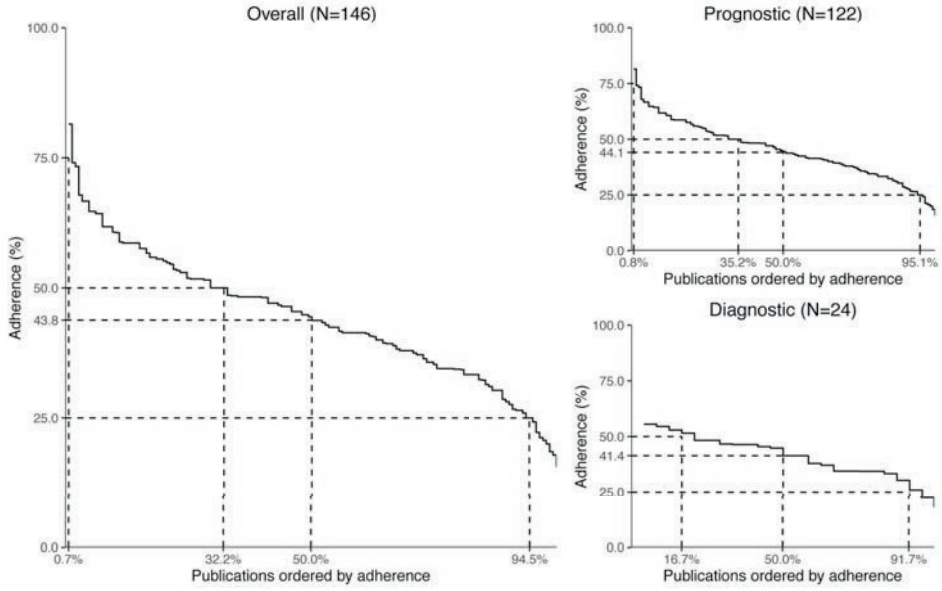


Figure 2: Reporting across publications: adherence to items of the TRIPOD statement

Reporting of individual TRIPOD items

Six TRIPOD items were reported in 75% or more of the 170 models, and 10 items in less than 25% (Table 1). Completeness of reporting of individual TRIPOD items is presented in Figure 3 and Supplemental Table 3 over all 170 models, and per type of prediction model study. The most notable findings for each section of the TRIPOD statement (title and abstract, introduction, methods, results, discussion, and other information) are described below.

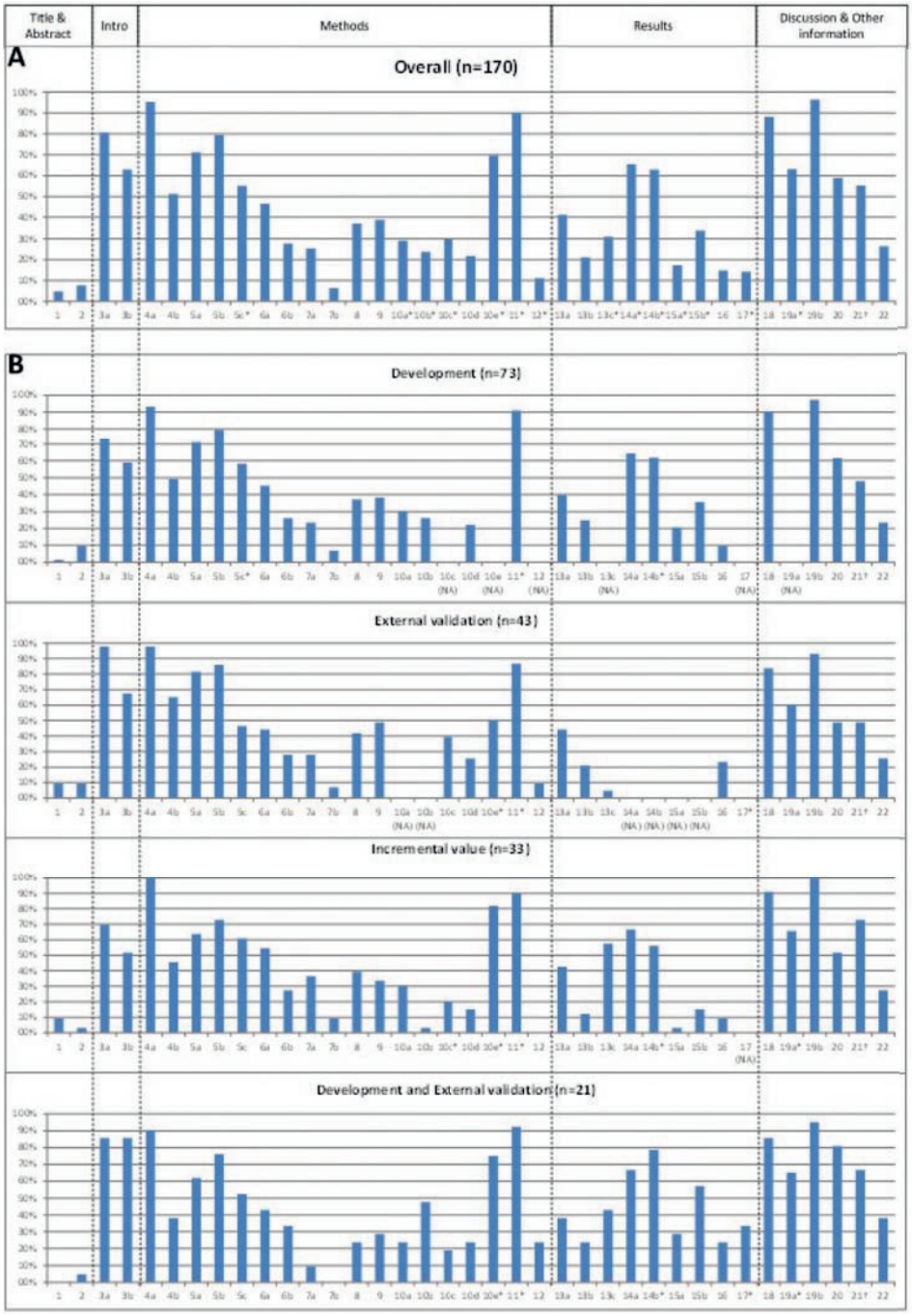
Table 1: Completeness of reporting of individual TRIPOD items (n=170 models).

Complete reporting for >75% of the models			Complete reporting for <25% of the models		
High reporting quality of TRIPOD items		%	Poor reporting quality of TRIPOD items		%
19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	96	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	24
4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	95	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	22
11	Provide details on how risk groups were created, if done.	90	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	21
18	Discuss any limitations of the study (such as non-representative sample, few events per predictor, missing data).	88	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	17
3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models	81	16	Report performance measures (with CIs) for the prediction model.	15

Table 1: Continued

Complete reporting for >75% of the models			Complete reporting for <25% of the models		
High reporting quality of TRIPOD items		%	Poor reporting quality of TRIPOD items		%
5b	Describe eligibility criteria for participants.	79	17	If done, report the results from any model updating (i.e., model specification, model performance).	14
			12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	11
			2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	8
			7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	6
			1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	5

Figure 3: (right page) Reporting of the items of the TRIPOD statement overall (A), and per type of prediction model study (B) (see Box 1 for list of items of the TRIPOD statement) NA: not applicable (not all items of the TRIPOD statement are relevant to all types of prediction model studies). Percentages are based on number of models for which an item was applicable (and thus should have been reported). *Where this number deviates from the total number of models, this is indicated. This concerns the following items (N=number of models for which the item was applicable): Overall: 5c (N=169), 10a (N=127), 10b (N=127), 10c (N=84), 10e (N=23), 11 (N=70), 12 (N=81), 13c (N=97), 14a (N=127), 14b (N=94), 15a (N=127), 15b (N=127), 17 (N=7), 19a (N=92) Development: 5c (N=72), 11 (N=22), 14b (N=55); External validation: 10e (N=8), 11 (N=15), 17 (N=4); Incremental value: 10c (N=20), 10e (N=11), 11 (N=20), 12 (N=17), 14b (N=25), 19a (N=29); Development and external validation: 10e (N=4), 11 (N=13), 14b (N=14), 17 (N=3), 19a (N=20). †Item 21 "Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets": the number of models for which this item was applicable is unknown. It probably was applicable to all models that reported this item. Instead of presenting a percentage of 100, we based the percentage on the total number of models.


Title & Abstract
Intro
Methods
Results
Discussion & Other information
A
Overall (n=170)
B
Development (n=73)
External validation (n=43)
Incremental value (n=33)
Development and External validation (n=21)

Title and abstract (items 1 and 2)

According to the TRIPOD statement, an informative title contains (synonyms for) the term *risk prediction model*, the *type of prediction model* study (i.e. development, external validation, incremental value, or combination), the *target population,* and *outcome to be predicted.* Eight of the 170 models (5%) addressed all four elements. The description of the type of prediction model study was the least reported element (12%). Depending on the type of prediction model study, complete reporting of abstracts required information for up to 12 elements. Thirteen of the models (8%) fulfilled all the requirements.

Introduction (item 3)

For 81% of the models complete information about background and rationale was provided (item 3a) and in 63% reporting of study objectives (item 3b), including a specification of the type of prediction model study, was considered complete.

Methods (items 4 – 12)

Source of data (item 4a; 95% reported) and eligibility criteria (item 5b; 79%) were among the best reported items for all four types of prediction model studies. Actions to blind assessment of (non-objective) outcomes (item 6b; 28%) and predictors (item 7b; 7%) were less well reported. Detailed predictor definitions (item 7a) were provided for 25% of the models. Also information about how missing data were handled (item 9) was incomplete for the majority of models (reported in 39%). Most aspects of statistical analysis were inadequately reported as well. How predictors were handled (item 10a) was described in 29% of the models. Model building procedures (item 10b) were specified in 24% overall, and particularly poor in incremental value reports (3%). Few studies (22%) described both discrimination and calibration as measures of model performance (item 10d).

Results (items 13 – 17)

Characteristics of participants (item 13b, complete reporting in 21%) were often reported without information regarding missing data for predictors and outcome. Two (5%) of the external validations presented demographics, distribution of predictors, and outcomes alongside those of the original development study (item 13c) and in combined reports of development and external validation this was done in 43%. The final model was presented in full (item 15a) in 17% of the models. For many models the intercept (or the cumulative baseline hazard (or baseline survival) for at least one time point in the case of survival models) was not provided. A small number of models provided information on both discrimination and calibration when reporting model performance (item 16; 15%). Discrimination was more frequently reported (79%) than calibration (29%).

Discussion (items 18 – 20)

An overall interpretation of the results (item 19b) was given for almost all included models of all types of prediction model studies (97%). The potential for clinical use and implications for future research (item 20) were discussed in 59% of the models.

Other information (items 21 and 22)

Information about the availability of supplementary resources (item 21) was provided in 55% of the models. Complete information regarding funding (item 22) was reported in 27%.

Discussion

Complete and accurate reporting of prediction model studies is required to critically appraise, externally validate, evaluate their impact, and eventually use prediction models in clinical practice. Our study shows that, regardless of the type of prediction model study and whether diagnostic or prognostic, more than half of the items deemed essential to report in prediction model publications according to the TRIPOD statement were not completely reported.

Highly problematic TRIPOD items in terms of reporting were items regarding title and abstract. These items, for which complete reporting requires information on multiple elements, were adequately reported for less than 10% of the models. In addition, details of study methods, especially blinding of outcome and predictor assessments, were provided for only a minority of reported models. Furthermore, information on follow-up, predictor definitions, model building procedures and handling of missing data were often lacking. Notable findings regarding the reporting of study results were that in over 70% of the included models the final model was not presented in enough detail to make predictions for new patients, and that the reporting of model performance was often incomplete. Items of the TRIPOD statement that were generally well reported addressed the source of data and eligibility criteria, risk groups (if applicable), study limitations, and overall interpretation of results.

Comparison with other studies

Our main finding of inadequate reporting in the majority of publications within 37 clinical domains is comparable to the findings of systematic reviews of prediction model studies performed in general medicine or specific clinical domains.[6-11] Inadequate reporting is considered to be a form of research waste.[18,19] Therefore, for many study types reporting guidelines were published in the last 20 years, such as the CONSORT (Consolidated Standards of Reporting Trials) statement in 1996 (updates in 2001 and 2010), the STARD

(Standards for Reporting of Diagnostic Accuracy) statement in 2003 (update in 2015), and REMARK (Reporting recommendations for tumour marker prognostic studies) in 2005.[20-24] Completeness of reporting before the introduction of these reporting guidelines was similar to our result of 44% adherence. Moher and colleagues (2001) evaluated 97 reports of randomized trials before the introduction of CONSORT and found adequate reporting for just over half of the items (58%).[25] In a systematic review of 16 studies evaluating the adherence to STARD, overall, 51% of items were adequately reported.[26] For six included studies with quantitative data before publication of STARD a range of 44% to 61% adherence was reported. An assessment of the reporting of prognostic studies of tumour markers was done shortly after the introduction of REMARK.[27,28] Ten (out of 20) items were evaluated, and, overall, articles adhered to 53% of these.

Strengths and limitations of this study

With this literature review we cover a broad literature base by including three major types of prediction model studies, both prognostic and diagnostic, across 37 clinical domains. Despite the use of a validated search strategy, we may have missed publications on prediction models. It is likely that the completeness of reporting of prediction models in these studies would have been worse. Furthermore, we selected studies from high impact journals. Therefore, our results on the completeness of reporting might be an optimistic representation of the reporting of prediction model studies in general.

We were strict in scoring adherence by requiring complete information on all elements of a TRIPOD item, e.g. complete reporting of model performance required the provision of both discrimination and calibration measures. This is in line with the nature of TRIPOD as having essential items needed to appraise and utilize a prediction model.

However, authors might have good reasons not to provide specific details regarding an item. For example, if they believe that their model should not be validated or used in clinical practice, they may have decided not to present the coefficients of the full model. In the current study we would have scored TRIPOD item 15a as “incompletely reported”. Although strict scoring potentially leads to poorer adherence results, it is needed for reasons of consistency.

We used two different denominators in our analyses, the number of publications (n=146) and the number of models (n=170). It implies that in the “model” analysis a number of publications were included multiple times. It is likely that results from the same publication although based on the reporting of different models are correlated. Given the descriptive nature of our analysis, we did not adjust for such a possible correlation.

We present results from studies that were published almost four years ago, nevertheless we expect these findings to be still applicable and relevant to current publications of prediction models. From evaluations of other reporting guidelines, like CONSORT and STARD, we know that it takes time to demonstrate the impact of a

reporting guideline on completeness of reporting and changes over several years might be small.[25,26,28-33] To our opinion, therefore, it is too early for a before-after comparison at this moment, and the focus should be on optimal implementation of TRIPOD first.

Implications for practice and areas for future research

Inadequate reporting impedes the use of all available evidence regarding a prediction model. First, as title and abstract were among the least well reported items, identifying publications of prediction model studies might be challenging. In addition, we found the reporting of model development often insufficiently detailed, which makes external validation almost impossible. As a consequence, a new model might be developed, rather than making use of an existing model. Also, without model specifications it is impossible to use the model in clinical practice. Finally, inadequate reporting hinders critical appraisal and, by that, the possibility of methodological investigation of sources of variation and bias in prediction model studies.

Experiences from other research areas indicate that the improvement in reporting after the introduction of a guideline is often slow and might be subtle.[25,26,28-33] Improving the completeness of reporting of prediction models is probably even more challenging, as it is a relatively young, less well known research field, with methodology in development and not yet strongly embedded in education. Moreover, the multivariable nature of prediction model studies and their focus on absolute probabilities rather than on comparative measures require the reporting of many details on methods and results. It should also be taken into account that practical issues, like word limits or journal requirements, could act as barriers for complete reporting.

The introduction of the TRIPOD statement was the first step in improving the reporting of prediction model studies. However, more activities should be undertaken to enhance the implementation of the TRIPOD statement. Active implementation involves a collaborative effort of developers of a reporting guideline and other stakeholders within the academic community, like journal editors and educational institutions. Apart from raising awareness and providing training, possible post-publication activities that are recommended are encouraging guideline endorsement, asking for feedback, and evaluating the impact of the reporting guideline.[34]

By highlighting the flaws in the reporting of prediction model studies, our results enable a targeted implementation strategy for the TRIPOD statement. Possible future activities are the development of educational materials and training regarding specific aspects of the reporting of prediction model studies. The examples of both adequate and suboptimal reporting within our dataset can be used in the training of different stakeholders. An initiative that already has been started by the TRIPOD Group is the development of specific guidance on informative reporting of prediction model studies in abstracts.[35] Furthermore, as TRIPOD is periodically being reappraised and will be updated if necessary, our study will provide useful input for modifications of specific TRIPOD items, related to either content, phrasing or more detailed explanation.[12] Finally,

our study will serve as a baseline measurement for future studies evaluating the impact of the introduction the TRIPOD statement.

Conclusion

Prediction models are poorly reported: more than half of the items that are considered essential for transparent reporting of a prediction model were not or were inadequately reported, especially with regard to details of the title, abstract, blinding, model building procedures, the final model, and model performance. The results of this study can be used to further develop and refine the implementation and increase the impact of the TRIPOD statement.

Acknowledgements

We thank René Spijker for performing the search for this comprehensive literature survey, and Daan Michels for his assistance with data extraction.

References

1. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10(2):e1001381.
2. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:b605.
3. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;98(9):691-8.
4. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012;98(9):683-90.
5. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ* 2009;338:b604.
6. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med* 2011;9:103.
7. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9(5):1-12.
8. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013;66(3):268-77.
9. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;14:40.
10. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;353:i2416.
11. Wen Z, Guo Y, Xu B, Xiao K, Peng T, Peng M. Developing Risk Prediction Models for Postoperative Pancreatic Fistula: a Systematic Review of Methodology and Reporting Quality. *Indian J Surg* 2016;78(2):136-43.
12. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162(1):55-63.
13. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.
14. 2012 Journal Citation Reports® Science Edition ed: Clarivate Analytics 2017.

15. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc* 2001;8(4):391-7.
16. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ* 2009;338:b375.
17. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med* 2013;10(2):e1001380.
18. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *Lancet* 2009;374(9683):86-9.
19. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 2014;383(9913):267-76.
20. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c869.
21. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med* 2010;152(11):726-32.
22. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015;351:h5527.
23. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;6(11):e012799.
24. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. REporting recommendations for tumour MARKer prognostic studies (REMARK). *Br J Cancer* 2005;93(4):387-91.
25. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA* 2001;285(15):1992-5.
26. Korevaar DA, van Enst WA, Spijker R, Bossuyt PM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. *Evid Based Med* 2014;19(2):47-54.
27. Mallett S, Timmer A, Sauerbrei W, Altman DG. Reporting of prognostic studies of tumour markers: a review of published articles in relation to REMARK guidelines. *Br J Cancer* 2010;102(1):173-80.
28. Sekula P, Mallett S, Altman DG, Sauerbrei W. Did the reporting of prognostic studies of tumour markers improve since the introduction of REMARK guideline? A comparison of reporting in published articles. *PLoS One* 2017;12(6):e0178531.

29. Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. *BMJ* 2010;340:c723.
30. Korevaar DA, Wang J, van Enst WA, Leeflang MM, Hooft L, Smidt N, et al. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. *Radiology* 2015;274(3):781-9.
31. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, et al. The quality of diagnostic accuracy studies since the STARD statement: has it improved? *Neurology* 2006;67(5):792-7.
32. Turner L, Shamseer L, Altman DG, Weeks L, Peters J, Kober T, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database Syst Rev* 2012;11:Mr000030.
33. Chan AW, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet* 2005;365(9465):1159-62.
34. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS Med* 2010;7(2):e1000217.
35. Reporting of clinical prediction model studies in journal and conference abstracts: TRIPOD for Abstracts. 24th Cochrane Colloquium; 2016 23-27 Oct; 2016; Seoul, Korea. John Wiley & Sons.

Supplemental material

Supplemental Table 1: Ten journals with the highest Journal Impact Factor within each of 37 categories (clinical domains) (2012 Journal Citation Reports® [Clarivate Analytics, 2017]) Full journal titles indicated with an * were included in more than one category.

Category (clinical domain)	Full journal title	Journal Impact Factor
Allergy	Journal of Allergy and Clinical Immunology*	12.047
	Allergy	5.883
	Clinical Reviews in Allergy & Immunology	5.590
	Clinical and Experimental Allergy	4.789
	Annals of Allergy Asthma & Immunology	3.449
	Current Opinion In Allergy and Clinical Immunology	3.398
	Pediatric Allergy and Immunology*	3.376
	Contact Dermatitis*	2.925
	Current Allergy and Asthma Reports	2.746
	Allergy Asthma & Immunology Research	2.653
Anesthesiology	Pain	5.644
	Anesthesiology	5.163
	British Journal of Anaesthesia	4.237
	Anaesthesia	3.486
	Regional Anesthesia and Pain Medicine	3.464
	Anesthesia and Analgesia	3.300
	European Journal of Pain	3.067
	Minerva Anestesiologica*	2.818
	European Journal of Anaesthesiology	2.792
	Pain Practice	2.605
Cardiac and cardiovasuclar systems	Circulation*	15.202
	European Heart Journal	14.097
	Journal of the American College of Cardiology	14.086
	Circulation Research*	11.861
	Nature Reviews Cardiology	10.400
	Circulation-Cardiovascular Genetics	6.728
	Circulation-Heart Failure	6.684
	Jacc-Cardiovascular Interventions	6.552
	Circulation-Cardiovascular Interventions	6.543
	Jacc-Cardiovascular Imaging*	6.164

Supplemental Table 1: Continued

Category (clinical domain)	Full journal title	Journal Impact Factor
Clinical neurology	Lancet Neurology	23.917
	Nature Reviews Neurology	15.518
	Alzheimers & Dementia	14.483
	Annals of Neurology	11.193
	Brain	9.915
	Acta Neuropathologica	9.734
	Sleep Medicine Reviews	8.681
	Neurology	8.249
	Archives of Neurology	7.685
	Neuro-Oncology	6.180
Critical care medicine	American Journal of Respiratory and Critical Care Medicine*	11.041
	Critical Care Medicine	6.124
	Chest*	5.854
	Intensive Care Medicine	5.258
	Critical Care	4.718
	Journal of Neurotrauma	4.295
	Resuscitation*	4.104
	Neurocritical Care	3.038
	Current Opinion In Critical Care	2.967
	Minerva Anestesiologica*	2.818
Dentistry. Oral surgery & medicine	Periodontology 2000	4.012
	Journal of Dental Research	3.826
	Clinical Implant Dentistry and Related Research	3.821
	Dental Materials	3.773
	Journal of Clinical Periodontology	3.688
	Clinical Oral Implants Research	3.433
	Journal of Dentistry	3.200
	Journal of Endodontics	2.929
	International Journal of Oral Science	2.719
	British Journal of Oral & Maxillofacial Surgery	2.717

Supplemental Table 1: Continued

Category (clinical domain)	Full journal title	Journal Impact Factor
Dermatology	Journal of Investigative Dermatology	6.193
	Pigment Cell & Melanoma Research	5.839
	Journal of the American Academy of Dermatology	4.906
	Archives of Dermatology	4.792
	British Journal of Dermatology	3.759
	Experimental Dermatology	3.578
	Journal of Dermatological Science	3.520
	Acta Dermato-Venereologica	3.487
	Contact Dermatitis*	2.925
	Skin Pharmacology and Physiology	2.885
Emergency medicine	Annals of Emergency Medicine	4.285
	Resuscitation*	4.104
	Emergencias	2.578
	Journal of Trauma-Injury Infection and Critical Care	2.348
	Injury-International Journal of the Care of the Injured	2.174
	Prehospital Emergency Care	1.859
	Academic Emergency Medicine	1.757
	American Journal of Emergency Medicine	1.704
	Scandinavian Journal of Trauma Resuscitation & Emergency Medicine	1.680
	Emergency Medicine Journal	1.645
Endocrinology & Metabolism	Endocrine Reviews	14.873
	Cell Metabolism	14.619
	Nature Reviews Endocrinology	11.025
	Trends In Endocrinology and Metabolism	8.901
	Frontiers In Neuroendocrinology	7.985
	Diabetes	7.895
	Diabetes Care	7.735
	Journal of Mammary Gland Biology and Neoplasia	7.524
	Journal of Pineal Research	7.304
	Antioxidants & Redox Signaling	7.189

Supplemental Table 1: Continued

Category (clinical domain)	Full journal title	Journal Impact Factor
Gastroenterology & Hepatology	Gastroenterology	12.821
	Hepatology	12.003
	Gut	10.732
	Nature Reviews Gastroenterology & Hepatology	10.426
	Journal of Hepatology	9.858
	Seminars In Liver Disease	8.274
	American Journal of Gastroenterology	7.553
	Clinical Gastroenterology and Hepatology	6.648
	Endoscopy*	5.735
	Gastrointestinal Endoscopy	5.210
Geriatrics & Gerontology	Neurobiology of Aging	6.166
	Ageing Research Reviews	5.953
	Aging Cell	5.705
	Journal of the American Medical Directors Association	5.302
	Frontiers In Aging Neuroscience	5.224
	Journals of Gerontology Series A-Biological Sciences and Medical Sciences	4.314
	American Journal of Geriatric Psychiatry	4.131
	Age	4.084
	Journal of the American Geriatrics Society	3.978
	Experimental Gerontology	3.911
Hematology	Circulation Research*	11.861
	Leukemia*	10.164
	Blood	9.060
	Stem Cells	7.701
	Arteriosclerosis Thrombosis and Vascular Biology*	6.338
	Thrombosis and Haemostasis*	6.094
	Journal of Thrombosis and Haemostasis*	6.081
	Blood Reviews	6.000
	Haematologica-the Hematology Journal	5.935
	Journal of Cerebral Blood Flow and Metabolism	5.398

Supplemental Table 1: Continued

Category (clinical domain)	Full journal title	Journal Impact Factor
Immunology	Annual Review of Immunology	36.556
	Nature Reviews Immunology	33.129
	Nature Immunology	26.199
	Immunity	19.795
	Journal of Experimental Medicine	13.214
	Immunological Reviews	12.155
	Journal of Allergy and Clinical Immunology*	12.047
	Trends In Immunology	9.486
	Clinical Infectious Diseases*	9.374
	Current Opinion In Immunology	8.771
Infectious diseases	Lancet Infectious Diseases	19.966
	Clinical Infectious Diseases*	9.374
	Aids	6.407
	Emerging Infectious Diseases	5.993
	Journal of Infectious Diseases	5.848
	Eurosurveillance	5.491
	Journal of Antimicrobial Chemotherapy	5.338
	Current Opinion In Infectious Diseases	4.870
	Current Opinion In Hiv and Aids	4.704
	Jaids-Journal of Acquired Immune Deficiency Syndromes	4.653
Integrative & complementary medicine	Alternative Medicine Review	4.857
	Phytomedicine	2.972
	Journal of Ethnopharmacology	2.755
	Integrative Cancer therapies	2.354
	American Journal of Chinese Medicine	2.281
	Complementary therapies In Medicine	2.093
	Bmc Complementary and Alternative Medicine	2.082
	Evidence-Based Complementary and Alternative Medicine	1.722
	Journal of Manipulative and Physiological therapeutics	1.647
	Journal of Alternative and Complementary Medicine	1.464

Supplemental Table 1: Continued

Category (clinical domain)	Full journal title	Journal Impact Factor
Medical laboratory technology	Clinical Chemistry	7.149
	Critical Reviews In Clinical Laboratory Sciences	3.783
	Advances In Clinical Chemistry	3.674
	Translational Research	3.490
	Clinical Chemistry and Laboratory Medicine	3.009
	Clinica Chimica Acta	2.850
	Archives of Pathology & Laboratory Medicine	2.781
	Clinical Biochemistry	2.450
	Therapeutic Drug Monitoring	2.234
	Cytometry Part B-Clinical Cytometry	2.231
Medicine. general & internal	New England Journal of Medicine	51.658
	Lancet	39.060
	Jama-Journal of the American Medical Association	29.978
	British Medical Journal	17.215
	Plos Medicine	15.253
	Annals of Internal Medicine	13.976
	Archives of Internal Medicine	10.579
	Bmc Medicine	6.679
	Canadian Medical Association Journal	6.465
	Journal of Internal Medicine	6.455
Obstetrics & Gynecology	Human Reproduction Update*	8.847
	Obstetrics and Gynecology	4.798
	Human Reproduction*	4.670
	Fertility and Sterility*	4.174
	Gynecologic Oncology	3.929
	American Journal of Obstetrics and Gynecology	3.877
	Bjog-An International Journal of Obstetrics and Gynaecology	3.760
	Ultrasound In Obstetrics & Gynecology	3.557
	Seminars In Reproductive Medicine*	3.211
	Menopause-the Journal of the North American Menopause Society	3.163

Supplemental Table 1: Continued

Category (clinical domain)	Full journal title	Journal Impact Factor
Oncology	Ca-A Cancer Journal For Clinicians	153.459
	Nature Reviews Cancer	35.000
	Lancet Oncology	25.117
	Cancer Cell	24.755
	Journal of Clinical Oncology	18.038
	Nature Reviews Clinical Oncology	15.031
	Jnci-Journal of the National Cancer Institute	14.336
	Leukemia*	10.164
	Cancer Discovery	10.143
	Biochimica Et Biophysica Acta-Reviews On Cancer	9.033
Ophthalmology	Progress In Retinal and Eye Research	9.439
	Ophthalmology	5.563
	Archives of Ophthalmology	3.826
	American Journal of Ophthalmology	3.631
	Investigative Ophthalmology & Visual Science	3.441
	Experimental Eye Research	3.026
	Survey of Ophthalmology	2.859
	Retina-the Journal of Retinal and Vitreous Diseases	2.825
	British Journal of Ophthalmology	2.725
	Ocular Surface	2.643
Orthopedics	American Journal of Sports Medicine*	4.439
	Osteoarthritis and Cartilage*	4.262
	Journal of Bone and Joint Surgery-American Volume	3.234
	Spine Journal	3.220
	Arthroscopy-the Journal of Arthroscopic and Related Surgery	3.103
	Journal of Orthopaedic & Sports Physical therapy*	2.947
	Journal of Orthopaedic Research	2.875
	Clinical Orthopaedics and Related Research	2.787
	Physical therapy*	2.778
	Acta Orthopaedica	2.736

Supplemental Table 1: Continued

Category (clinical domain)	Full journal title	Journal Impact Factor
Otorhinolaryngology	Ear and Hearing	3.262
	Jaro-Journal of the Association For Research In Otolaryngology	2.952
	Head and Neck-Journal For the Sciences and Specialties of the Head and Neck	2.833
	Hearing Research	2.537
	Audiology and Neuro-Otology	2.318
	Otology & Neurotology	2.014
	Laryngoscope	1.979
	Dysphagia	1.938
	Clinical Otolaryngology	1.869
	Archives of Otolaryngology-Head & Neck Surgery	1.779
Pediatrics	Journal of the American Academy of Child and Adolescent Psychiatry*	6.970
	Pediatrics	5.119
	Archives of Pediatrics & Adolescent Medicine	4.282
	Journal of Pediatrics	4.035
	European Child & Adolescent Psychiatry	3.699
	Pediatric Infectious Disease Journal	3.569
	Seminars In Fetal & Neonatal Medicine	3.505
	Archives of Disease In Childhood-Fetal and Neonatal Edition	3.451
	Pediatric Allergy and Immunology*	3.376
	Archives of Disease In Childhood	3.051

Supplemental Table 1: Continued

Category (clinical domain)	Full journal title	Journal Impact Factor
Peripheral vascular disease	Circulation*	15.202
	Circulation Research*	11.861
	Hypertension	6.873
	Arteriosclerosis Thrombosis and Vascular Biology*	6.338
	Stroke	6.158
	Thrombosis and Haemostasis*	6.094
	Journal of Thrombosis and Haemostasis*	6.081
	Current Opinion In Lipidology	5.839
	Atherosclerosis Supplements	4.333
	Seminars In Thrombosis and Hemostasis	4.216
Primary health care	Annals of Family Medicine	4.613
	Primary Care Respiratory Journal	2.191
	British Journal of General Practice	2.034
	Scandinavian Journal of Primary Health Care	1.905
	Family Practice	1.828
	Canadian Family Physician	1.808
	Journal of the American Board of Family Medicine	1.758
	American Family Physician	1.611
	Bmc Family Practice	1.609
	Primary Care Diabetes	1.609
Psychiatry	Molecular Psychiatry	14.897
	American Journal of Psychiatry	14.721
	Archives of General Psychiatry	13.772
	Biological Psychiatry	9.247
	World Psychiatry	8.974
	Neuropsychopharmacology	8.678
	Schizophrenia Bulletin	8.486
	Psychotherapy and Psychosomatics	7.230
	Journal of the American Academy of Child and Adolescent Psychiatry*	6.970
	British Journal of Psychiatry	6.606

Supplemental Table 1: Continued

Category (clinical domain)	Full journal title	Journal Impact Factor
Public. Environmental and Occupational health	Epidemiologic Reviews	9.269
	Environmental Health Perspectives	7.260
	International Journal of Epidemiology	6.982
	Who Technical Report Series	6.100
	Epidemiology	5.738
	Journal of Clinical Epidemiology	5.332
	Bulletin of the World Health Organization	5.250
	European Journal of Epidemiology	5.118
	American Journal of Epidemiology	4.780
	Cancer Epidemiology Biomarkers & Prevention	4.559
Radiology. Nuclear medicine and Medical imaging	Human Brain Mapping	6.878
	Radiology	6.339
	Neuroimage	6.252
	Jacc-Cardiovascular Imaging*	6.164
	Circulation-Cardiovascular Imaging	5.795
	Journal of Nuclear Medicine	5.774
	Investigative Radiology	5.460
	European Journal of Nuclear Medicine and Molecular Imaging	5.114
	International Journal of Radiation Oncology Biology Physics	4.524
	Radiotherapy and Oncology	4.520
Rehabilitation	Journal of Head Trauma Rehabilitation	4.443
	Neurorehabilitation and Neural Repair	4.278
	Ieee Transactions On Neural Systems and Rehabilitation Engineering	3.255
	Journal of Orthopaedic & Sports Physical therapy*	2.947
	Physical therapy*	2.778
	Supportive Care In Cancer	2.649
	Journal of Neuroengineering and Rehabilitation	2.567
	American Journal of Speech-Language Pathology	2.448
	Archives of Physical Medicine and Rehabilitation	2.358
	Journal of Physiotherapy	2.255

Supplemental Table 1: Continued

Category (clinical domain)	Full journal title	Journal Impact Factor
Reproductive biology	Human Reproduction Update*	8.847
	Human Reproduction*	4.670
	Molecular Human Reproduction	4.542
	Fertility and Sterility*	4.174
	Biology of Reproduction	4.027
	Reproduction	3.555
	American Journal of Reproductive Immunology	3.317
	Seminars In Reproductive Medicine*	3.211
	Reproductive Toxicology	3.141
	Placenta	3.117
Respiratory system	American Journal of Respiratory and Critical Care Medicine*	11.041
	Thorax	8.376
	European Respiratory Journal	6.355
	Chest*	5.854
	Journal of Heart and Lung Transplantation*	5.112
	Journal of Thoracic Oncology	4.473
	American Journal of Respiratory Cell and Molecular Biology	4.148
	Respiratory Research	3.642
	Journal of Thoracic and Cardiovascular Surgery	3.526
	American Journal of Physiology-Lung Cellular and Molecular Physiology	3.523
Rheumatology	Nature Reviews Rheumatology	9.745
	Annals of the Rheumatic Diseases	9.111
	Arthritis and Rheumatism	7.477
	Current Opinion In Rheumatology	5.191
	Arthritis Research & therapy	4.302
	Osteoarthritis and Cartilage*	4.262
	Rheumatology	4.212
	Seminars In Arthritis and Rheumatism	3.806
	Arthritis Care & Research	3.731
	Best Practice & Research In Clinical Rheumatology	3.550

Supplemental Table 1: Continued

Category (clinical domain)	Full journal title	Journal Impact Factor
Sport sciences	Exercise Immunology Review	7.053
	Exercise and Sport Sciences Reviews	5.283
	Sports Medicine	5.237
	Medicine and Science In Sports and Exercise	4.475
	American Journal of Sports Medicine*	4.439
	British Journal of Sports Medicine	3.668
	Journal of Applied Physiology	3.484
	Scandinavian Journal of Medicine & Science In Sports	3.214
	Journal of Orthopaedic & Sports Physical therapy*	2.947
	Journal of Science and Medicine In Sport	2.899
Surgery	Annals of Surgery	6.329
	American Journal of Transplantation*	6.192
	Endoscopy*	5.735
	Journal of Neurology Neurosurgery and Psychiatry	4.924
	American Journal of Surgical Pathology	4.868
	British Journal of Surgery	4.839
	Journal of the American College of Surgeons	4.500
	Surgery For Obesity and Related Diseases	4.121
	Annals of Surgical Oncology	4.120
	Archives of Surgery	4.100
Transplantation	American Journal of Transplantation*	6.192
	Journal of Heart and Lung Transplantation*	5.112
	Stem Cells and Development	4.670
	Cell Transplantation	4.422
	Liver Transplantation	3.944
	Biology of Blood and Marrow Transplantation	3.940
	Transplantation	3.781
	Bone Marrow Transplantation	3.541
	Nephrology Dialysis Transplantation	3.371
	Current Opinion In Organ Transplantation	3.272

Supplemental Table 1: Continued

Category (clinical domain)	Full journal title	Journal Impact Factor
Tropical medicine	Plos Neglected Tropical Diseases	4.569
	Malaria Journal	3.400
	Tropical Medicine & International Health	2.938
	Acta Tropica	2.787
	American Journal of Tropical Medicine and Hygiene	2.534
	Transactions of the Royal Society of Tropical Medicine and Hygiene	1.823
	Memorias Do Instituto Oswaldo Cruz	1.363
	Annals of Tropical Medicine and Parasitology	1.313
	Journal of Vector Borne Diseases	1.041
	Journal of Tropical Pediatrics	1.006
Urology & Nephrology	European Urology	10.476
	Journal of the American Society of Nephrology	8.987
	Nature Reviews Nephrology	7.943
	Kidney International	7.916
	American Journal of Kidney Diseases	5.294
	Clinical Journal of the American Society of Nephrology	5.068
	Nature Reviews Urology	4.793
	Current Opinion In Nephrology and Hypertension	3.964
	Prostate	3.843
	Journal of Urology	3.696

Supplemental Table 2: Pubmed search strategy on July 4th 2014

	Hits
((Validat*[tiab] OR Predict*[ti] OR Rule*[tiab]) OR (Predict*[tiab] AND (Outcome*[tiab] OR Risk*[tiab] OR Model*[tiab])) OR ((History[tiab] OR Variable*[tiab] OR Criteria[tiab] OR Scor*[tiab] OR Characteristic*[tiab] OR Finding*[tiab] OR Factor*[tiab]) AND (Predict*[tiab] OR Model*[tiab] OR Decision*[tiab] OR Identif*[tiab] OR Prognos*[tiab])) OR (Decision*[tiab] AND (Model*[tiab] OR Clinical*[tiab] OR logistic models[mesh])) OR (Prognostic[tiab] AND (History[tiab] OR Variable*[tiab] OR Criteria[tiab] OR Scor*[tiab] OR Characteristic*[tiab] OR Finding*[tiab] OR Factor*[tiab] OR Model*[tiab]))) AND (0091-6749[is] OR 0105-4538[is] OR 1080-0549[is] OR 0954-7894[is] OR 1081-1206[is] OR 1528-4050[is] OR 0905-6157[is] OR 0105-1873[is] OR 1529-7322[is] OR 2092-7355[is] OR 0304-3959[is] OR 0003-3022[is] OR 0007-0912[is] OR 0003-2409[is] OR 1098-7339[is] OR 0003-2999[is] OR 1090-3801[is] OR 0375-9393[is] OR 0265-0215[is] OR 1530-7085[is] OR 0009-7322[is] OR 0195-668X[is] OR 0735-1097[is] OR 0009-7330[is] OR 1759-5002[is] OR 1942-325X[is] OR 1941-3289[is] OR 1936-8798[is] OR 1941-7640[is] OR 1936-878X[is] OR 1474-4422[is] OR 1759-4758[is] OR 1552-5260[is] OR 0364-5134[is] OR 0006-8950[is] OR 0001-6322[is] OR 1087-0792[is] OR 0028-3878[is] OR 0003-9942[is] OR 1522-8517[is] OR 1073-449X[is] OR 0090-3493[is] OR 0012-3692[is] OR 0342-4642[is] OR 1466-609X[is] OR 0897-7151[is] OR 0300-9572[is] OR 1541-6933[is] OR 1070-5295[is] OR 0375-9393[is] OR 0906-6713[is] OR 0022-0345[is] OR 1523-0899[is] OR 0109-5641[is] OR 0303-6979[is] OR 0905-7161[is] OR 0300-5712[is] OR 0099-2399[is] OR 1674-2818[is] OR 0266-4356[is] OR 0022-202X[is] OR 1755-1471[is] OR 0190-9622[is] OR 0003-987X[is] OR 0007-0963[is] OR 0906-6705[is] OR 0923-1811[is] OR 0001-5555[is] OR 0105-1873[is] OR 1660-5527[is] OR 0196-0644[is] OR 0300-9572[is] OR 1137-6821[is] OR 0022-5282[is] OR 0020-1383[is] OR 1090-3127[is] OR 1069-6563[is] OR 0735-6757[is] OR 1757-7241[is] OR 1472-0205[is] OR 0163-769X[is] OR 1550-4131[is] OR 1759-5029[is] OR 1043-2760[is] OR 0091-3022[is] OR 0012-1797[is] OR 0149-5992[is] OR 1083-3021[is] OR 0742-3098[is] OR 1523-0864[is] OR 0016-5085[is] OR 0270-9139[is] OR 0017-5749[is] OR 1759-5045[is] OR 0168-8278[is] OR 0272-8087[is] OR 0002-9270[is] OR 1542-3565[is] OR 0013-726X[is] OR 0016-5107[is] OR 0028-4793[is] OR 0140-6736[is] OR 0098-7484[is] OR 1756-1833[is] OR 1549-1676[is] OR 0003-4819[is] OR 0003-9926[is] OR 1741-7015[is] OR 0820-3946[is] OR 0954-6820[is] OR 0197-4580[is] OR 1568-1637[is] OR 1474-9718[is] OR 1525-8610[is] OR 1663-4365[is] OR 1079-5006[is] OR 1064-7481[is] OR 0161-9152[is] OR 0002-8614[is] OR 0531-5565[is] OR 0009-7330[is] OR 0887-6924[is] OR 0006-4971[is] OR 1066-5099[is] OR 1079-5642[is] OR 0340-6245[is] OR 1538-7933[is] OR 0268-960X[is] OR 0390-6078[is] OR 0271-678X[is] OR 0732-0582[is] OR 1474-1733[is] OR 1529-2908[is] OR 1074-7613[is] OR 0022-1007[is] OR 0105-2896[is] OR 0091-6749[is] OR 1471-4906[is] OR 1058-4838[is] OR 0952-7915[is] OR 1473-3099[is] OR 1058-4838[is] OR 0269-9370[is] OR 1080-6040[is] OR 0022-1899[is] OR 1560-7917[is] OR 0305-7453[is] OR 0951-7375[is] OR 1746-630X[is] OR 1525-4135[is] OR 1089-5159[is] OR 0944-7113[is] OR 0378-8741[is] OR 1534-7354[is] OR 0192-415X[is] OR 0965-2299[is] OR 1472-6882[is] OR 1741-427X[is] OR 0161-4754[is] OR 1075-5535[is] OR 0009-9147[is] OR 1040-8363[is] OR 0065-2423[is] OR 1931-5244[is] OR 1434-6621[is] OR 0009-8981[is] OR 0003-9985[is] OR 0009-9120[is] OR 0163-4356[is] OR 1552-4949[is] OR 1355-4786[is] OR 0029-7844[is] OR 0268-1161[is] OR 0015-0282[is] OR 0090-8258[is] OR 0002-9378[is] OR 1470-0328[is] OR 0960-7692[is] OR 1526-8004[is] OR 1072-3714[is] OR 0007-9235[is] OR 1474-175X[is] OR 1470-2045[is] OR 1535-6108[is] OR 0732-183X[is] OR 1759-4774[is] OR 0027-8874[is] OR 0887-6924[is] OR 2159-8274[is] OR 1350-9462[is] OR 0161-6420[is] OR 0003-9950[is] OR 0002-9394[is] OR 0146-0404[is] OR 0014-4835[is] OR 0039-6257[is] OR 0275-004X[is] OR 0007-1161[is] OR 1542-0124[is] OR 0363-5465[is] OR 1063-4584[is] OR 0021-9355[is] OR 1529-	4871

Supplemental Table 2: Continued

9430[is] OR 0749-8063[is] OR 0190-6011[is] OR 0736-0266[is] OR 0009-921X[is] OR 0031-9023[is] OR 1745-3674[is] OR 0196-0202[is] OR 1525-3961[is] OR 1043-3074[is] OR 0378-5955[is] OR 1420-3030[is] OR 1531-7129[is] OR 0023-852X[is] OR 0179-051X[is] OR 1749-4478[is] OR 0886-4470[is] OR 0890-8567[is] OR 0031-4005[is] OR 1072-4710[is] OR 0022-3476[is] OR 1018-8827[is] OR 0891-3668[is] OR 1744-165X[is] OR 1359-2998[is] OR 0905-6157[is] OR 0003-9888[is] OR 0009-7322[is] OR 0009-7330[is] OR 0194-911X[is] OR 1079-5642[is] OR 0039-2499[is] OR 0340-6245[is] OR 1538-7933[is] OR 0957-9672[is] OR 1567-5688[is] OR 0094-6176[is] OR 1544-1709[is] OR 1471-4418[is] OR 0960-1643[is] OR 0281-3432[is] OR 0263-2136[is] OR 0008-350X[is] OR 1557-2625[is] OR 0002-838X[is] OR 1471-2296[is] OR 1751-9918[is] OR 1359-4184[is] OR 0002-953X[is] OR 0003-990X[is] OR 0006-3223[is] OR 1723-8617[is] OR 0893-133X[is] OR 0586-7614[is] OR 0033-3190[is] OR 0890-8567[is] OR 0007-1250[is] OR 0193-936X[is] OR 0091-6765[is] OR 0300-5771[is] OR 0512-3054[is] OR 1044-3983[is] OR 0895-4356[is] OR 0042-9686[is] OR 0393-2990[is] OR 0002-9262[is] OR 1055-9965[is] OR 1065-9471[is] OR 0033-8419[is] OR 1053-8119[is] OR 1936-878X[is] OR 1941-9651[is] OR 0161-5505[is] OR 0020-9996[is] OR 1619-7070[is] OR 0360-3016[is] OR 0167-8140[is] OR 0885-9701[is] OR 1545-9683[is] OR 1534-4320[is] OR 0190-6011[is] OR 0031-9023[is] OR 0941-4355[is] OR 1743-0003[is] OR 1058-0360[is] OR 0003-9993[is] OR 1836-9553[is] OR 1355-4786[is] OR 0268-1161[is] OR 1360-9947[is] OR 0015-0282[is] OR 0006-3363[is] OR 1470-1626[is] OR 1046-7408[is] OR 1526-8004[is] OR 0890-6238[is] OR 0143-4004[is] OR 1073-449X[is] OR 0040-6376[is] OR 0903-1936[is] OR 0012-3692[is] OR 1053-2498[is] OR 1556-0864[is] OR 1044-1549[is] OR 1465-993X[is] OR 0022-5223[is] OR 1040-0605[is] OR 1759-4790[is] OR 0003-4967[is] OR 0004-3591[is] OR 1040-8711[is] OR 1478-6354[is] OR 1063-4584[is] OR 1462-0324[is] OR 0049-0172[is] OR 2151-464X[is] OR 1521-6942[is] OR 1077-5552[is] OR 0091-6331[is] OR 0112-1642[is] OR 0195-9131[is] OR 0363-5465[is] OR 0306-3674[is] OR 8750-7587[is] OR 0905-7188[is] OR 0190-6011[is] OR 1440-2440[is] OR 0003-4932[is] OR 1600-6135[is] OR 0013-726X[is] OR 0022-3050[is] OR 0147-5185[is] OR 0007-1323[is] OR 1072-7515[is] OR 1550-7289[is] OR 1068-9265[is] OR 0004-0010[is] OR 1600-6135[is] OR 1053-2498[is] OR 1547-3287[is] OR 0963-6897[is] OR 1527-6465[is] OR 1083-8791[is] OR 0041-1337[is] OR 0268-3369[is] OR 0931-0509[is] OR 1087-2418[is] OR 1935-2735[is] OR 1475-2875[is] OR 1360-2276[is] OR 0001-706X[is] OR 0002-9637[is] OR 0035-9203[is] OR 0074-0276[is] OR 0003-4983[is] OR 0972-9062[is] OR 0142-6338[is] OR 0302-2838[is] OR 1046-6673[is] OR 1759-5061[is] OR 0085-2538[is] OR 0272-6386[is] OR 1555-9041[is] OR 1759-4812[is] OR 1062-4821[is] OR 0270-4137[is] OR 0022-5347[is]) AND (2014/05/01 : 2014/06/01[dp])

Supplemental table 3: Reporting of the items of the TRIPOD statement

	Development	External validation	Incremental value	Development and external validation	Overall
	N=73	N=43	N=33	N=21	N=170
Items of the TRIPOD statement	n (%)	n (%)	n (%)	n (%)	n (%)
TITLE AND ABSTRACT					
1. Title: identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1 (1)	4 (9)	3 (9)	0 (0)	8 (5)
2. Abstract: provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	7 (10)	4 (9)	1 (3)	1 (5)	13 (8)
INTRODUCTION					
3. Background and objectives:					
a. Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	54 (74)	42 (98)	23 (70)	18 (86)	137 (81)
b. Specify the objectives, including whether the study describes the development or validation of the model or both.	43 (59)	29 (67)	17 (52)	18 (86)	107 (63)

Supplemental table 3: Continued

	Development	External validation	Incremental value	Development and external validation	Overall
	N=73	N=43	N=33	N=21	N=170
Items of the TRIPOD statement	n (%)	n (%)	n (%)	n (%)	n (%)
METHODS					
4. Source of data:					
a. Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	68 (93)	42 (98)	33 (100)	19 (91)	162 (95)
b. Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	36 (49)	28 (65)	15 (46)	8 (38)	87 (51)
5. Participants:					
a. Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	52 (71)	35 (81)	21 (64)	13 (62)	121 (71)
b. Describe eligibility criteria for participants.	58 (80)	37 (86)	24 (73)	16 (76)	135 (79)
c. Give details of treatments received, if relevant.	42/72 (58)*	20 (47)	20 (61)	11 (52)	93/169 (55)*
6. Outcome:					
a. Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	33 (45)	19 (44)	18 (55)	9 (43)	79 (47)

Supplemental table 3: Continued

	Development	External validation	Incremental value	Development and external validation	Overall
	N=73	N=43	N=33	N=21	N=170
Items of the TRIPOD statement	n (%)	n (%)	n (%)	n (%)	n (%)
b. Report any actions to blind assessment of the outcome to be predicted.	19 (26)	12 (28)	9 (27)	7 (33)	47 (28)
7. Predictors:					
a. Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	17 (23)	12 (28)	12 (36)	2 (10)	43 (25)
b. Report any actions to blind assessment of predictors for the outcome and other predictors.	5 (7)	3 (7)	3 (9)	0 (0)	11 (7)
8. Sample size: explain how the study size was arrived at.	27 (37)	18 (42)	13 (39)	5 (24)	63 (37)
9. Missing data: Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	28 (38)	21 (49)	11 (33)	6 (29)	66 (39)
10. Statistical analysis methods:					
a. Describe how predictors were handled in the analyses.	22 (30)	NA	10 (30)	5 (24)	37/127 (29)*
b. Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	19 (26)	NA	1 (3)	10 (48)	30/127 (24)*

Supplemental table 3: Continued

	Development	External validation	Incremental value	Development and external validation	Overall
	N=73	N=43	N=33	N=21	N=170
Items of the TRIPOD statement	n (%)	n (%)	n (%)	n (%)	n (%)
c. For validation, describe how the predictions were calculated.	NA	17 (40)	4/20 (20)*	4 (19)	25/84 (30)*
d. Specify all measures used to assess model performance and, if relevant, to compare multiple models.	16 (22)	11 (26)	5 (15)	5 (24)	37 (22)
e. Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA	4/8 (50)*	9/11 (82)*	3/4 (75)*	16/23 (70)*
11. Risk groups: Provide details on how risk groups were created, if done.	20/22 (91)*	13/15 (87)*	18/20 (90)*	12/13 (92)*	63/70 (90)*
12. Development vs. validation: for validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	NA	4 (9)	0/17 (0)*	5 (24)	9/81 (11)*
RESULTS					
13. Participants:					
a. Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	29 (40)	19 (44)	14 (42)	8 (38)	70 (41)

Supplemental table 3: Continued

	Development	External validation	Incremental value	Development and external validation	Overall
	N=73	N=43	N=33	N=21	N=170
Items of the TRIPOD statement	n (%)	n (%)	n (%)	n (%)	n (%)
b. Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	18 (25)	9 (21)	4 (12)	5 (24)	36 (21)
c. For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	NA	2 (5)	19 (58)	9 (43)	30/97 (31)*
14. Model development:					
a. Specify the number of participants and outcome events in each analysis.	47 (64)	NA	22 (67)	14 (67)	83/127 (65)*
b. If done, report the unadjusted association between each candidate predictor and outcome.	34/55 (62)*	NA	14/25 (56)*	11/14 (79)*	59/94 (63)*
15. Model specification:					
a. Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	15 (21)	NA	1 (3)	6 (29)	22/127 (17)*
b. Explain how to the use the prediction model.	26 (36)	NA	5 (15)	12 (57)	43/127 (34)*

Supplemental table 3: Continued

	Development	External validation	Incremental value	Development and external validation	Overall
	N=73	N=43	N=33	N=21	N=170
Items of the TRIPOD statement	n (%)	n (%)	n (%)	n (%)	n (%)
16. Model performance: report performance measures (with CIs) for the prediction model.	7 (10)	10 (23)	3 (9)	5 (24)	25 (15)
17. Model-updating: if done, report the results from any model updating (i.e., model specification, model performance).	NA	0/4 (0)*	NA	1/3 (33)*	1/7 (14)*
DISCUSSION					
18. Limitations: discuss any limitations of the study (such as non-representative sample, few events per predictor, missing data).	66 (90)	36 (84)	30 (91)	18 (86)	150 (88)
19. Interpretation:					
a. For validation, discuss the results with reference to performance in the development data, and any other validation data.	NA	26 (61)	19/29 (66)*	13/20 (65)*	58/92 (63)*
b. Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	71 (97)	40 (93)	33 (100)	20 (95)	164 (97)
20. Implications: discuss the potential clinical use of the model and implications for future research.	45 (62)	21 (49)	17 (52)	17 (81)	100 (59)

Supplemental table 3: Continued

	Development	External validation	Incremental value	Development and external validation	Overall
	N=73	N=43	N=33	N=21	N=170
Items of the TRIPOD statement	n (%)	n (%)	n (%)	n (%)	n (%)
OTHER INFORMATION					
21. Supplementary information: provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	35 (48)†	21 (49)†	24 (73)†	14 (67)†	94 (55)†
22. Funding: give the source of funding and the role of the funders for the present study.	17 (23)	11 (26)	9 (27)	8 (38)	45 (27)

Chapter 8

Empirical evidence on the impact of study characteristics on the performance of prediction models: a meta-epidemiological study

Johanna AAG Damen
Thomas PA Debray
Romin Pajouheshnia
Johannes B Reitsma
Rob JPM Scholten
Karel GM Moons
Lotty Hooft

Manuscript in preparation

Abstract

Background: Meta-epidemiological studies have shown that shortcomings in study design can lead to biased estimates of treatment effects and diagnostic test accuracy. So far, it remains unclear to what extent study characteristics affect estimates of prognostic model performance.

Objectives: To empirically assess the relation between study characteristics and the results of external validation studies of multivariable prognostic models.

Methods: We searched electronic databases for systematic reviews of prognostic models published between 2010 and 2016. Reviews from non-overlapping clinical fields were selected if they reported common performance measures (either the concordance (c)-statistic or the ratio of observed over expected number of events (OE ratio)) from ten or more validations of the same prognostic model. From the included external validation studies we extracted study design features, population characteristics, methods of predictor and outcome assessment, the handling of missing data and the aforementioned performance measures. Random effects meta-regression was used to quantify the association between the characteristics of validation studies and model performance.

Results: We included 10 systematic reviews, describing a total of 224 external validations, of which 221 reported c-statistics and 124 OE ratios. Associations between study characteristics and model performance were heterogeneous across systematic reviews. C-statistics were most associated with population characteristics and predictor and outcome measurement, e.g. validation in a continent different from the development study resulted in a higher c-statistic, compared to validation in the same continent (difference in logit c-statistic 0.10 [95% CI 0.04, 0.16]), and validations with eligibility criteria comparable to the development study were associated with higher c-statistics compared to narrower criteria (difference in logit c-statistic 0.21 [95% CI 0.07, 0.35]). Using a case-control design was associated with higher OE ratios, compared to using data from a cohort (difference in log OE ratio 0.97 [95% CI 0.38, 1.55]).

Conclusions: Variation in performance of prognostic models across studies or settings is mainly associated with variation in case-mix, study designs, and predictor and outcome measurement methods. Researchers developing and validating prognostic models should realise the influence of these study characteristics on the predictive performance of prognostic models.

Introduction

Prediction models, including diagnostic and prognostic models, estimate the probability that an individual has or will develop a certain outcome (e.g. disease or complication). Hereto, they combine multiple predictors into an estimate of an individual's risk.[1] Before using a prediction model in clinical practice it is recommended to validate the performance of the model in a population other than the population in which the model was developed (so called external validation studies).[2] Such studies assess whether model predictions remain sufficiently accurate across different settings and populations. Obviously, it is important that the methodological quality of external validation studies is good, as otherwise estimates of the prediction model's performance may be biased and thereby lead to misleading conclusions on its generalizability.

Systematic reviews have found that the performance of existing prediction models often varies substantially across external validation studies of those models.[3-5] These differences may not only appear when validation studies are small (due to random variation), but may also arise when model predictions are invalid because the model is applied in very different populations (e.g. the association between predictors in the model and the outcome are different) or when design-related characteristics of the validation study (e.g. measurement methods or variable definitions) are not well aligned with the original development study.[2,6]

To provide empirical evidence of the association of study characteristics with prediction model performance, a meta-epidemiological approach can be used. Studies using this approach have shown the influence of study characteristics on the effectiveness of interventions studied in randomized trials and on the accuracy of diagnostic tests.[7-12] For diagnostic prediction models some evidence exists that suggests estimates of performance may be biased in studies with certain study characteristics. One study found a higher diagnostic odds ratio in case-control studies, studies with differential outcome verification (i.e. using different outcome assessments across study individuals), and with low sample size.[13] To date, no meta-epidemiological study has been performed investigating the possible impact of study characteristics on measures of the predictive performance of a prognostic model upon external validation, which is commonly quantified in terms of discrimination and calibration.[14]

The aim of this study was to investigate sources of heterogeneity in the predictive performance of prognostic models. A meta-epidemiological approach was used to synthesize evidence from a range of clinical fields. This study can serve as empirical evidence for design and analysis related bias in prognostic model studies.

Methods

Search and selection of systematic reviews

We used an existing database (last updated on March 27, 2017) consisting of studies evaluating multiple existing prediction models, including narrative or systematic reviews of prediction models, or head-to-head comparisons of multiple prediction models validated on a specific dataset (See Supplement 1 for details of the search strategy). To construct this database, references identified by the search were screened for eligibility by one reviewer (GSC) on title, abstract and, if necessary, on full text. Subsequently, the full text of all articles in the database were screened for eligibility to the current project by another reviewer (JAAGD). We selected systematic reviews of prognostic models (i.e. diagnostic models were excluded) that included at least ten studies that externally validated the same prognostic model, and that presented the performance of these models in terms of discrimination (concordance (c)-statistic or area under the receiver operating characteristic (AUC) curve), or calibration (observed expected (OE) ratio). Discrimination is the ability of the model to distinguish between people who will and who will not develop the outcome of interest, while calibration reflects the overall agreement between the total number of observed and predicted ('expected') events.[14] We excluded systematic reviews that selected studies based on specific study characteristics (e.g. we excluded systematic reviews that did not include primary studies with a sample size below 100, if we were not able to identify the primary studies that had been excluded for this reason). Furthermore, we excluded reviews of prognostic models in which the weights of predictors in the original model were based on expert opinion rather than on coefficients estimated from a formal statistical approach. If more than one systematic review on the same prognostic model was identified, we included the one with the broadest inclusion criteria (e.g. reviews focussing on specific patient populations were not preferred if a review with a broader population was available), or the most recent review, or the one with the highest number of external validations (in this order of preference). When multiple prognostic models for the same condition were described in a systematic review which all fulfilled the selection criteria, we included the model with the highest number of external validations.

Selection of the primary external validations from the included systematic reviews

From the included systematic reviews we collected the primary studies in which the prediction models were developed and externally validated. For primary studies for which no measure of discrimination (c-statistic) or calibration (OE ratio) was reported in the systematic review, we checked the full text of the primary external validation study, and if performance was not reported, these studies were excluded.

If primary external validation studies described multiple external validations of the same model and if there was no overlap in included participants between these

external validations (e.g. a model was validated in two different cohorts, or a model was validated in men and women separately), data were extracted for every external validation separately. If a model was validated multiple times on the same population (described in either one or multiple publications), we selected the external validation that was included in the systematic review. If the systematic review included all those external validations, we selected the one in which the study population and predicted outcome most closely resembled the population and outcome of the original model.

Data extraction and preparation

We extracted relevant features of design and conduct according to existing checklists on quality assessment (CHARMS) and reporting of prediction model studies (TRIPOD).[15-17] Information about study characteristics of studies in which the models were developed were extracted from the corresponding development papers. Information about study characteristics of primary external validation studies were first extracted from systematic reviews. This information was subsequently checked using the external validation studies and, if necessary, additional information was extracted by one reviewer (JAAGD or RP). Items we extracted included study type (e.g. external validation only, development of a new model and external validation of a model), study design (e.g. existing cohort, existing RCT), dependency of investigators (validation by independent investigators or investigators also involved in the development study), eligibility criteria for participant inclusion, setting, location (continent), study dates, number of centers, follow-up time and prediction horizon, age and gender distribution, deletion or substitution of predictors, outcome definition and measurement method, sample size and number of events, handling of missing data, and model performance (see Supplement 2 for details). The data extraction form was piloted on multiple articles by all reviewers (JAAGD, TPAD, LH, KGMM, RP, JBR, RJPMS).

For analysis purposes, some study characteristics had to be categorized or transformed (Supplement 2). For example, eligibility criteria of the validation study as compared to the development study had to be judged and categorized as comparable, narrower (if subgroups included in the development study were excluded from the validation study), broader (if subgroups excluded from the development study were included in the validation study), mixture (a combination of the two), or unclear. For setting, location, predictors and outcome a similar categorization was used. If data on study characteristics were not reported in the primary external validation studies, these were either categorized as 'unclear' (in case of categorical study variables), or the study was excluded from the analyses of that (missing) study characteristic (in case of continuous study variables, such as sample size). In order to improve comparability between reviews, we standardized continuous study variables separately for each systematic review, i.e. for every variable we divided the mean by the standard deviation of all external validations identified from the same systematic review.

Statistical analyses

We used a two-staged approach to study the possible association between study characteristics and predictive performance.

In the first stage, we fitted a univariable meta-regression model for every study characteristic within each systematic review with the logit c-statistic or log OE ratio as outcome variable.[18] The regression coefficients estimated from this meta-regression model indicate the difference in logit c-statistic or log OE ratio between a certain category of a study characteristic and a chosen reference category (i.e. the category that was present in most systematic reviews) of that characteristic.

In the second stage, these regression coefficients were pooled by the use of a random effects model. This reflected the average influence of the study characteristic on model performance across all systematic reviews. For continuous characteristics, the regression coefficients obtained in the first stage were jointly pooled across reviews, using bivariate meta-analysis.[19,20] For categorical characteristics the results of univariable meta-analyses are presented.

We planned to perform multivariable analyses to assess the association between various study characteristics in combination and the performance of prediction models, but due to the paucity of data we were not able to do so. All analyses are described in more detail in Supplement 3.

Results

Identification and selection of studies

The search identified 2037 studies, of which 496 were included in the database and screened on full text, and 66 were further assessed (Figure 1). Finally, ten systematic reviews were included.[21-30] These reviews addressed external validations of the following prognostic models: ABCD2,[31] Essen Stroke Risk Score (ESRS),[32] EuroSCORE,[33] Framingham,[34] FRAX,[35] Injury Severity Score (ISS),[36] model for end-stage liver disease (MELD),[37] Pneumonia Severity Index (PSI),[38] Revised Cardiac Risk Index (RCRI),[39] and Simplified Acute Physiology Score (SAPS) 3[40] (Table 1). The reviews included 248 primary external validation studies with 274 external model validations (one study could describe multiple model validations). During data extraction, 73 of 274 validations were eventually excluded (most often for not reporting a performance measure), and 20 additional external model validations were identified (Figure 1). This resulted in the inclusion of 224 external validations, of which 221 could be included in the analyses of the c-statistic, and 124 in the analyses of the OE ratio. For the OE ratio, only validations of the EuroSCORE, Framingham, FRAX, PSI, RCRI and SAPS 3 prognostic models were included, due to the very low number of reported OE ratios in the validations studies for the other four prognostic models.

Description of included validations

The number of external validations within each systematic review ranged from 11 to 30 (Table 1), and the median (IQR) sample size and number of events were 1069 (418-3043), and 92 (36-248), respectively. Most studies used an existing registry (N=104, 46%), or existing cohort (N=74, 33%) to validate the prognostic model. The median (IQR) c-statistic and OE ratio were 0.73 (0.64-0.82), and 0.92 (0.64-1.26), respectively. Predictive performance of the models was highly heterogeneous, even for external validations of the same prediction model, as indicated by the wide prediction intervals (Table 1).

Not all information on the study characteristics was reported for all external validations (Table S1). Information was often unclear (e.g. for outcome definitions (N=83, 37%) and handling of missing data (N=105, 47%)) or missing (e.g. case-mix information such as mean age (N=28, 13%) and gender distribution (N=16, 7%)).

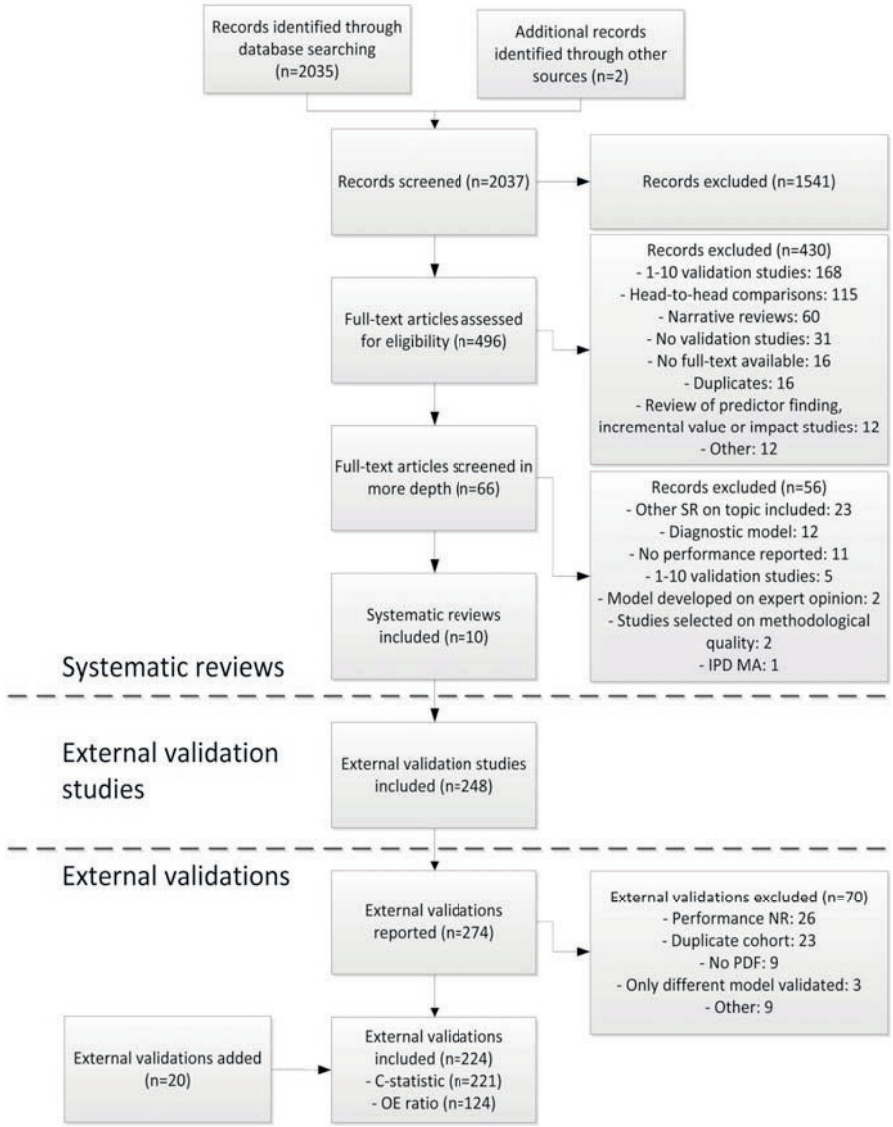


Figure 1: Flow chart of study selection.
SR: systematic review, IPD: individual participant data, MA: meta-analysis, NR: not reported, c: concordance, OE: observed expected.

Table 1: Description of included reviews and prediction models

Systematic review	Giles 2010[21]	Thompson 2014[22]	Siregar 2012[23]	Damen[30]	Marques 2015[24]	Tohira 2012[25]	Klein 2013[26]	Chalmers 2011[27]	Ford 2010[28]	Nassar 2014[29]
Model	ABCD2[31]	ESRS[32]	EuroSCORE[33]	Framingham[34]	FRAX[35]	ISS[36]	MELD[37]	PSI[38]	RCRI[39]	SAPS 3[40]
Population	Patients with TIA	Adults with a previous CVD event	Adult patients who underwent cardiac surgery under cardiopulmonary bypass	Men without previous CHD event	General population	Injured patients	Patients with liver cirrhosis but without hepatocellular carcinoma who underwent elective transjugular intrahepatic portosystemic shunts	Inpatients with community-acquired pneumonia	Patients aged >=50 years who underwent nonemergent noncardiac procedures	ICU patients
Geographical location	United States and UK	Canada, United States, Europe	Europe	United States	Europe, Canada, Japan, United States, UK, Australia	United States	United States	United States	United States	Worldwide
Patient recruitment	1981-1998	1992-1995	1995	1971-1974	1980-1999	1968-1969	1991-1995	1989	1989-1994	2002
Predicted outcome	Stroke	Recurrent ischaemic stroke, MI and vascular death	Mortality	CHD	Osteoporotic fractures	All-cause mortality	All-cause mortality	30-day hospital mortality	Major cardiac complications	Hospital mortality
Prediction horizon	2 days	1 year	30 days	10 years	10 years	3 months	3 months	30 days	1 year	90 days

Table 1: Continued

Systematic review	Giles 2010[21]	Thompson 2014[22]	Siregar 2012[23]	Damen[30]	Marques 2015[24]	Tohira 2012[25]	Klein 2013[26]	Chalmers 2011[27]	Ford 2010[28]	Nassar 2014[29]
Performance development study										
C-statistic	0.66 [95% CI 0.60-0.71]	NR	0.7875	0.74	0.63	NR	NR	0.84	0.759 (SE 0.032)	0.848
OE ratio	NR*	NR*	NR*	NR*	NR*	NR*	NR*	NR*	NR*	1.00 (95% CI 0.98-1.02)
***Pooled* performance validation studies**										
Number of external validations included in analyses	16	11	22	23	30	34	14	24	23	27
C-statistic [95% CI]	0.66 [0.61, 0.71]	0.60 [0.58, 0.62]	0.79 [0.77, 0.81]	0.68 [0.65, 0.71]	0.66 [0.63, 0.68]	0.86 [0.83, 0.88]	0.64 [0.59, 0.68]	0.80 [0.77, 0.82]	0.69 [0.65, 0.72]	0.83 [0.80, 0.85]
95% PI	[0.54, 0.77]	[0.57, 0.63]	[0.74, 0.83]	[0.56, 0.78]	[0.54, 0.76]	[0.62, 0.96]	[0.48, 0.77]	[0.64, 0.89]	[0.53, 0.81]	[0.66, 0.92]
OE ratio [95% CI]	NA	NA	0.54 [0.42, 0.68]	0.58 [0.45, 0.76]	1.10 [0.83, 1.47]	NA	NA	0.94 [0.83, 1.06]	2.70 [1.72, 4.25]	0.89 [0.77, 1.03]
95% PI	NA	NA	[0.19, 1.51]	[0.20, 1.74]	[0.31, 3.93]	NA	NA	[0.55, 1.60]	[0.35, 20.75]	[0.42, 1.91]

TIA: transient ischaemic attack, CVD: cardiovascular disease, CHD: coronary heart disease, ICU: intensive care unit, UK: United Kingdom, MI: myocardial infarction, NR: not reported, CI: confidence interval, PI: prediction interval, NA: not assessed. *As the models are optimally fit in the development dataset, all OE ratios should be close to 1.

Discrimination

Pooled models

The pooled analyses across all systematic reviews (Figure 2 and Figure S1) showed that validation in a continent different from the development study was associated with a higher c-statistic, compared to validation in the same continent, and multicenter versus single center validation studies were associated with a lower c-statistic. Comparable eligibility criteria for participant inclusion were also associated with higher c-statistics compared to narrower criteria, whereas a broader setting was associated with a lower c-statistic compared to a setting comparable to the development study. Although not statistically significant, validations with changes made to the predictors, or in which it was unclear whether all predictors were correctly measured, tended to have lower c-statistics compared to validations where no changes were made. In various reviews we found an association between the c-statistic and numerous other study characteristics, such as the study design, comparability of outcome definition, prediction horizon, sample size and number of events, and mean age of study participants (Figure 3, S1 and S2), only these were often not statistically significant when pooled together.

Variation across reviews

Across reviews we found effects of many study characteristics on the c-statistic although this was rather heterogeneous, and confidence intervals often overlapped (Figure 3 and Figure S2). For example, for study design, in six systematic reviews a higher c-statistic was found for validations that used an existing registry compared to an existing cohort, while in three reviews a lower c-statistic was found. In three systematic reviews we found a higher c-statistic in validations by independent investigators, while in five a lower c-statistic was found.

For other study characteristics, directions of associations were more consistent. For example, for most systematic reviews, validation studies with eligibility criteria narrower compared to the criteria used in the development study had a lower c-statistics while broader eligibility criteria were associated with higher c-statistics (Figure S2). C-statistics were also (slightly) higher in external validations with a setting comparable to the development study. Validation in a continent other than the development study in general was associated with a higher c-statistic, and multicenter studies had lower c-statistics compared to single center studies. External validations in which it was unclear if there were changes made to the predictors had lower c-statistics (Figure S2).

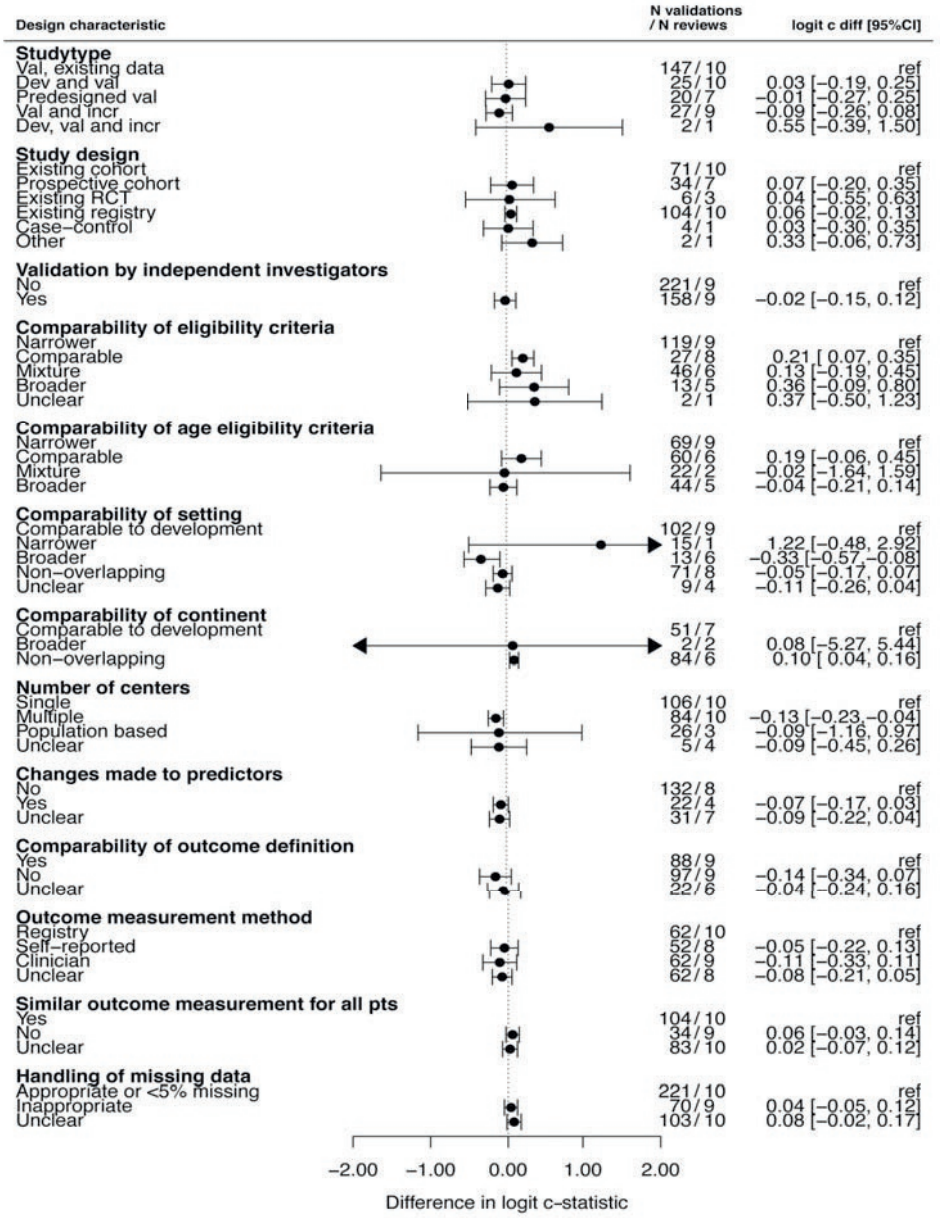


Figure 2: Influence of study characteristics on difference in logit c-statistic with regard to a reference category across 221 external validation studies and 10 different prediction models.

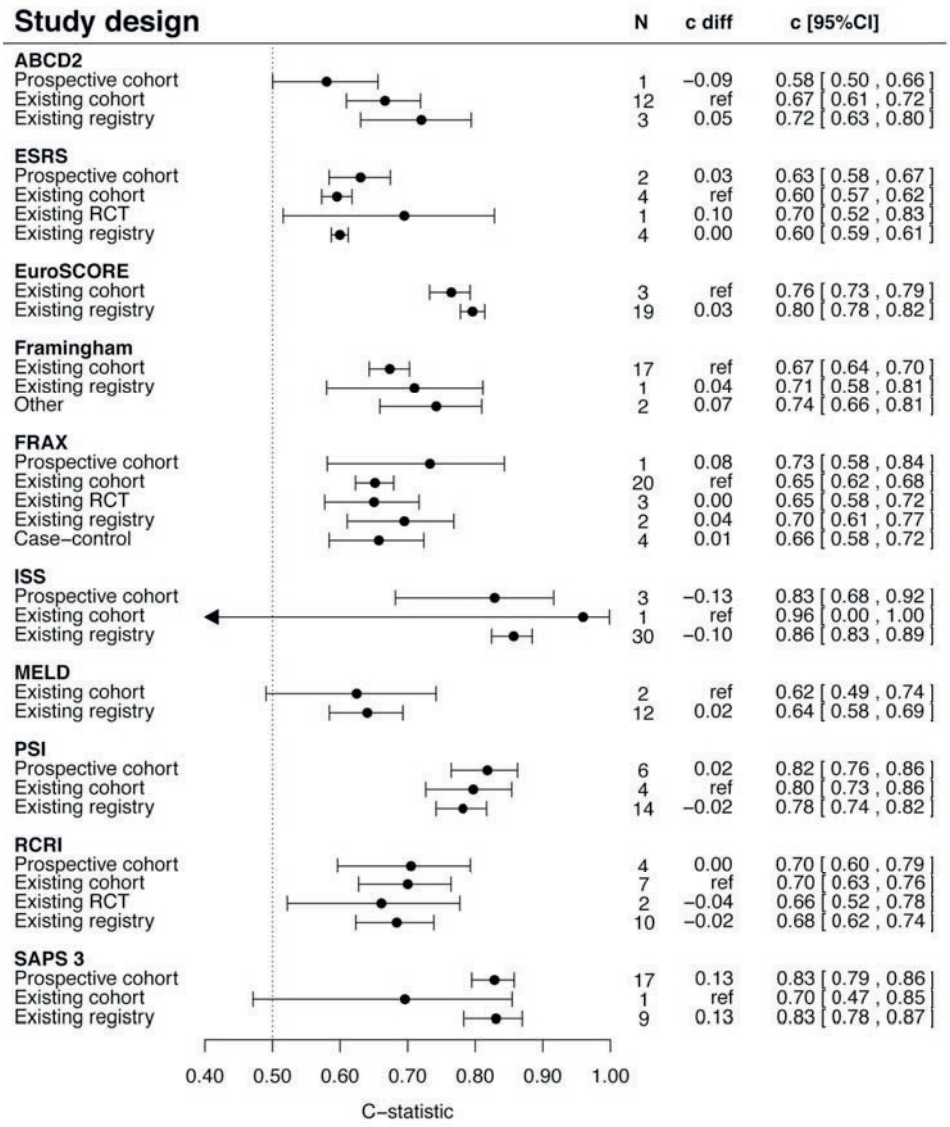


Figure 3: C-statistic for categories of study design, pooled using univariable meta-regression analyses within each systematic review. N represents the number of external validation studies in a specific category. C diff represents the difference in c-statistic with regard to a reference category (indicated with 'ref').

Calibration

Pooled analyses

We found a significant association between study design and the OE ratio (Figure 4); using data from a case-control study resulted in higher OE ratios, compared to using data from an existing cohort (though based on three external validations). Furthermore, higher OE ratios were found for studies in which the outcome was assessed by a panel of clinicians as compared to using a registry. In various reviews we found an association between the c-statistic and numerous other study characteristics, such as the duration of follow-up, year in which recruitment was started, sample size, standard deviation of age, and setting (Figure 4, S3 and S4), only these were often not statistically significant when pooled together.

Variation across reviews

For other categories of study design (other than the use of a case-control design), heterogeneous associations were found across systematic reviews (Figure 5). The associations of most other study characteristics with the OE ratio were also most often not consistent across systematic reviews (Figure S3 and S4). For example, for two systematic reviews external validations with appropriate handling of missing data had OE ratios closer to 1 compared to inappropriate handling of missing data, while in two reviews, OE ratios were further away from 1. Only for the continent in which the model was validated, directions were more consistent; OE ratios were closer to 1 if the continent was comparable to the development, compared to validations in different continents (Figure S4).

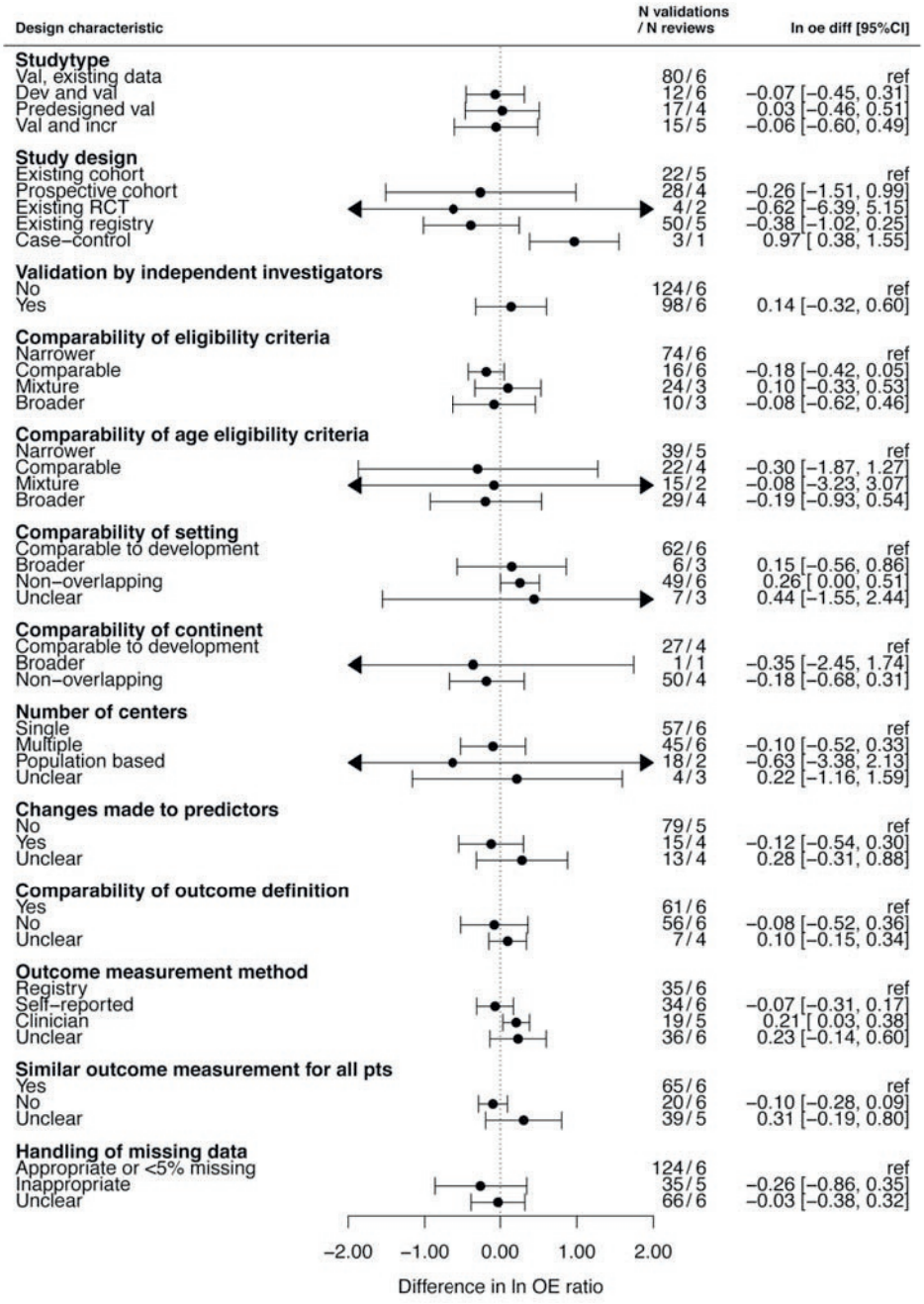


Figure 4: Influence of study characteristics on difference in ln OE ratio with regard to a reference category across 124 external validation studies and 6 different prediction models.

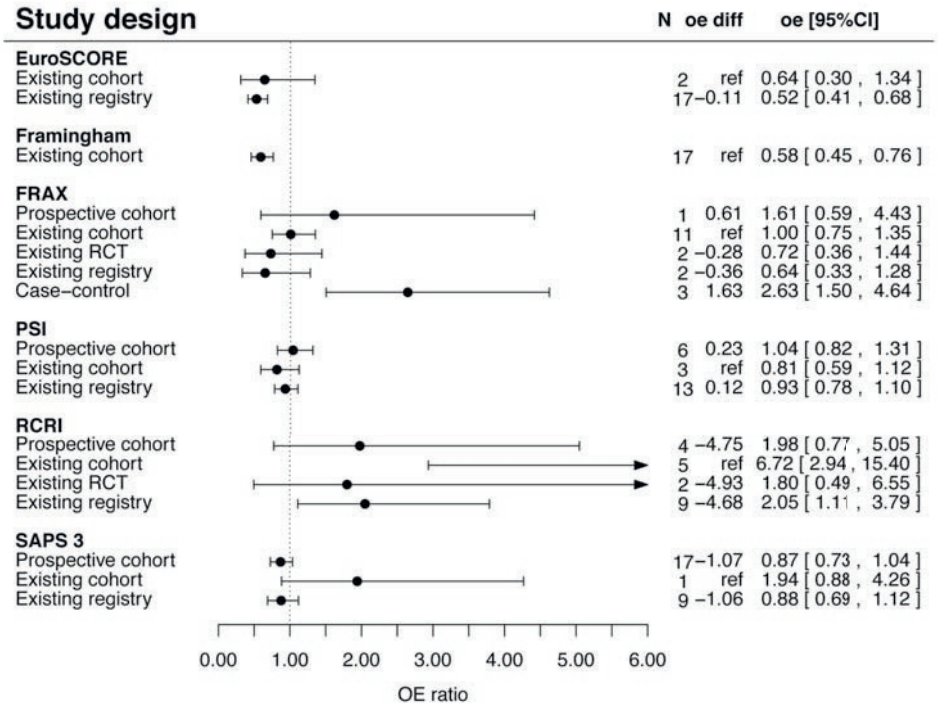


Figure 5: OE ratio for categories of study design, pooled using univariable meta-regression analyses within each systematic review. N represents the number of external validation studies in a specific category. OE diff represents the difference in OE ratio with regard to a reference category (indicated with 'ref').

Discussion

Summary of findings

Using a meta-analytical approach, we studied the association between study characteristics of external validation studies and the estimated performance of prognostic models across ten clinical domains. We focused on objective study characteristics that can be extracted from published reports. Unfortunately, reporting of the primary external validation studies was often inadequate. Key study characteristics, such as outcome definitions, handling of missing data, and even model calibration estimates were infrequently reported. Still, we found associations between various study characteristics and a model's predictive performance. Changes in a model's predictive performance were notably found in relation to validation studies with a case-control (versus cohort) design, with differences in case-mix, in continent (in which the model is validated), in eligibility criteria, in clinical setting, in number of centers (included in the validation study), and in differences in predictor and outcome assessments. For example, we noticed lower c-statistics in validations with unclear predictor measurement and validations that made changes to the predictors in the model (either deleted or substituted predictors).

Comparison to previous research

Our findings, i.e. the trends in the associations between study characteristics and model performance measures (though not always statistically significant), are in agreement with various previous simulation studies.[14,41-44] For example, we confirmed that studies with more variation in case-mix show higher c-statistics, and we noticed lower c-statistics when a predictor was omitted from the model. However, surprisingly, we found lower c-statistics in studies with a broader setting and when the number of centers in a study was higher.

We also found a higher OE ratio in studies with a case-control design. Both simulation studies and meta-epidemiological studies in the fields of diagnostic tests and (mainly diagnostic) prediction models, have shown biased effect measures in studies using a case-control design.[11-14] Further, we found that the OE ratio was influenced by the method of outcome assessment, in agreement with previous studies that showed that higher diagnostic odds ratios were found in studies with differential outcome verification.[13] We finally expected to find lower OE ratios when the validation population differed from the development population (e.g. in terms of case-mix).[14] We could not systematically confirm this across all reviews, likely caused by heterogeneity between systematic reviews as indicated by the wide confidence intervals. Finally, we could not fully confirm the association between sample size and model performance that was previously found, although we did find some trends in part of the reviews.[13]

Explanations, strengths and limitations

In this meta-epidemiological study we addressed associations between study characteristics and predictive performance of prognostic models, using data from more than 200 validations from 10 reviews. We expected more statistically significant associations between the predefined study characteristics and model performance across the systematic reviews. Although we included every systematic review that described at least ten external validation studies of the same prognostic model, our analyses appeared to still be hampered by relatively low numbers of external validations within each systematic review, combined with extreme heterogeneity within and across systematic reviews.

Conceptually, there are many potential sources of heterogeneity in prediction model performance, such as differences in population characteristics, predictor definitions, outcome definitions and in statistical analyses. These issues may act in isolation but more likely in combination, causing differences in model performance across systematic reviews and within systematic reviews. The combined effect of different study characteristics on the heterogeneity of prediction model performance across validation studies, is ideally addressed by adopting multivariable meta-regression models with the observed model performance estimates of the validation studies as dependent variable and the characteristics of multiple design features as independent variables. This multivariable meta-regression was unfortunately not feasible here due to the limited number of validation studies within the individual reviews.

A general limitation of all meta-epidemiological studies, is the possibility that the effect of a certain study characteristic differs across systematic reviews which may nullify the effect when pooled together. We found numerous conflicting associations between study characteristics and reported predictive performance measures that were not confirmed in the pooled analyses.

Also, it is possible that the effect caused by individual study characteristics is small and therefore difficult to detect. Moreover, there might be some misclassification of study characteristics, caused either by our misinterpretation of what is reported, or by a lack of reporting which could have diluted the effects of the study characteristics. In addition, the c-statistic is often considered to be an insensitive measure to quantify changes in model performance.[45-47] In previous simulation studies, the c-statistic and OE ratio appeared to be strongly influenced by case-mix differences,[14,41,48] which may mask the possible (smaller) effects from design-related characteristics. Other measures that are less sensitive to case-mix differences, such as the calibration slope, could, however, not be studied here simply because they are (almost) never reported.[3]

We found greater variation in the methods used by external validation studies between models than within validations of the same model. For example, multiple imputation is the preferred method for handling missing data in prediction modelling.[43,44] However, in the field of cardiovascular disease, all researchers handle missing data by

performing a complete case analysis, while in the field of mortality prediction in surgical patients, all researchers fill in 'normal' values if a value is missing. For the SAPS 3 model, most external validation studies used a prospective study design, while most external validation studies for the other models used existing datasets.

Finally, given the explorative nature of our analyses in order to generate further insight whether and when design features have an impact on the performance of prediction models, we did not statistically correct for multiple testing.

Implications for future research

In agreement with many previously conducted systematic reviews,[3,49-53] we found poor reporting of prediction model studies. Meta-epidemiological studies of prediction model studies would highly benefit from complete reporting according to the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement.[16,17]

We also believe that more research is urgently needed to evaluate under which circumstances certain design choices may lead to expected heterogeneity in prediction model performance, and to incorporate these issues in the appraisal of prediction modeling studies. There is a need for more guidance on how to score items of critical appraisal checklists for prediction modeling studies, such as the CHARMS checklist.[15]

Several options exist to gain more empirical insight in design related bias in prognostic studies. Firstly, meta-epidemiological researchers can collect more external validation studies and try to correct for all issues that cause variation in performance of a model. We believe, however, that this is currently not feasible as we already included every systematic reviews describing at least ten validations of the same prediction model. A second and much more efficient option is to collect the individual participant data (IPD) for all studies included in this review to directly study the effect of study characteristics on model performance.[54-56] Using IPD, it will also be possible to study different performance measures, like the case-mix adjusted c-statistic[41,57] and calibration slope.[14] Thirdly, new simulation studies could be performed to get more insight in design related bias. Researchers could for example study the effect of using a different outcome definition or prediction horizon on the c-statistic of a model.

Conclusion

In this meta-epidemiological study we found empirical evidence for an association between study characteristics and predictive performance of prognostic models. We found that predictive performance of prognostic models upon external validation is highly heterogeneous, but sensitive to various study characteristics, such as study design, case-mix, eligibility criteria, setting, and methods of predictor and outcome

measurement. It is important that these characteristics are thus emphasized in the reporting and appraisal of prediction model studies. However, for a large part the observed heterogeneity in model performance remained unexplained, which is likely caused by the high number of factors that cause heterogeneity in predictive performance and may act in opposite directions whereas a multivariable meta-regression analysis across reviews simply was not possible.

Acknowledgments

The authors would like to acknowledge Prof. Gary S Collins (GSC) for building the database with systematic reviews of prediction models, which served as a basis for this paper.

References

1. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012;98(9):683-90.
2. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2014.
3. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;353:i2416.
4. Alba AC, Agoritsas T, Jankowski M, Courvoisier D, Walter SD, Guyatt GH, et al. Risk prediction models for mortality in ambulatory patients with heart failure: a systematic review. *Circ Heart Fail* 2013;6(5):881-9.
5. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. *Breast Cancer Res Treat* 2012;132(2):365-77.
6. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ* 2009;338:b375.
7. Page MJ, Higgins JP, Clayton G, Sterne JA, Hrobjartsson A, Savovic J. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. *PLoS One* 2016;11(7):e0159267.
8. Berkman ND, Santaguida PL, Viswanathan M, Morton SC. AHRQ Methods for Effective Health Care. *The Empirical Evidence of Bias in Trials Measuring Treatment Differences*. Rockville (MD): Agency for Healthcare Research and Quality (US), 2014.
9. Savovic J, Jones H, Altman D, Harris R, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. *Health Technol Assess* 2012;16(35):1-82.
10. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 2008;336(7644):601-5.
11. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282(11):1061-6.
12. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174(4):469-76.
13. Ban JW, Emparanza JI, Urreta I, Burls A. Design Characteristics Influence Performance of Clinical Prediction Rules in Validation: A Meta-Epidemiological Study. *PLoS One* 2016;11(1):e0145779.

14. Steyerberg E. *Clinical prediction models: a practical approach to development, validation, and updating*: Springer Science & Business Media, 2008.
15. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11(10):e1001744.
16. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162(1):55-63.
17. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.
18. Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res* 2017:962280217705678.
19. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2015.
20. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;356:i6460.
21. Giles MF, Rothwell PM. Systematic review and pooled analysis of published and unpublished validations of the ABCD and ABCD2 transient ischemic attack risk scores. *Stroke* 2010;41(4):667-73.
22. Thompson DD, Murray GD, Dennis M, Sudlow CL, Whiteley WN. 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. *BMC Med* 2014;12:58.
23. Siregar S, Groenwold RH, de Heer F, Bots ML, van der Graaf Y, van Herwerden LA. Performance of the original EuroSCORE. *Eur J Cardiothorac Surg* 2012;41(4):746-54.
24. Marques A, Ferreira RJ, Santos E, Loza E, Carmona L, da Silva JA. The accuracy of osteoporotic fracture risk prediction tools: a systematic review and meta-analysis. *Ann Rheum Dis* 2015;74(11):1958-67.
25. Tohira H, Jacobs I, Mountain D, Gibson N, Yeo A. Systematic review of predictive performance of injury severity scoring tools. *Scand J Trauma Resusc Emerg Med* 2012;20:63.
26. Klein KB, Stafinski TD, Menon D. Predicting survival after liver transplantation based on pre-transplant MELD score: a systematic review of the literature. *PLoS One* 2013;8(12):e80661.

27. Chalmers JD, Mandal P, Singanayagam A, Akram AR, Choudhury G, Short PM, et al. Severity assessment tools to guide ICU admission in community-acquired pneumonia: systematic review and meta-analysis. *Intensive Care Med* 2011;37(9):1409-20.
28. Ford MK, Beattie WS, Wijeysundera DN. Systematic review: prediction of perioperative cardiac complications and mortality by the revised cardiac risk index. *Ann Intern Med* 2010;152(1):26-35.
29. Nassar AP, Malbouisson LM, Moreno R. Evaluation of Simplified Acute Physiology Score 3 performance: a systematic review of external validation studies. *Crit Care* 2014;18(3):R117.
30. Damen JA, Pajouheshnia R, Heus P, Moons KGM, Reitsma JB, Scholten RJPM, et al. Performance of the Framingham risk models and Pooled Cohort Equations for predicting 10-year risk of cardiovascular disease: a systematic review and meta-analysis (unpublished). Manuscript submitted for publication.
31. Rothwell PM, Giles MF, Flossmann E, Lovelock CE, Redgrave JN, Warlow CP, et al. A simple score (ABCD) to identify individuals at high early risk of stroke after transient ischaemic attack. *Lancet* 2005;366(9479):29-36.
32. Diener HC, Ringleb PA, Savi P. Clopidogrel for the secondary prevention of stroke. *Expert Opin Pharmacother* 2005;6(5):755-64.
33. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg* 1999;16(1):9-13.
34. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97(18):1837-47.
35. Kanis JA, Oden A, Johnell O, Johansson H, De Laet C, Brown J, et al. The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. *Osteoporos Int* 2007;18(8):1033-46.
36. Baker SP, O'Neill B, Haddon W, Jr., Long WB. The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. *J Trauma* 1974;14(3):187-96.
37. Malinchoc M, Kamath PS, Gordon FD, Peine CJ, Rank J, ter Borg PC. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology* 2000;31(4):864-71.
38. Fine MJ, Auble TE, Yealy DM, Hanusa BH, Weissfeld LA, Singer DE, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med* 1997;336(4):243-50.
39. Lee TH, Marcantonio ER, Mangione CM, Thomas EJ, Polanczyk CA, Cook EF, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation* 1999;100(10):1043-9.

40. Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P, Campos RA, et al. SAPS 3--From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* 2005;31(10):1345-55.
41. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010;172(8):971-80.
42. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol* 2012;12:82.
43. Held U, Kessels A, Garcia Aymerich J, Basagana X, Ter Riet G, Moons KG, et al. Methods for Handling Missing Variables in Risk Prediction Models. *Am J Epidemiol* 2016;184(7):545-51.
44. Janssen KJ, Vergouwe Y, Donders AR, Harrell FE, Jr., Chen Q, Grobbee DE, et al. Dealing with missing predictor values when applying clinical prediction models. *Clin Chem* 2009;55(5):994-1001.
45. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21(1):128-38.
46. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27(2):157-72; discussion 207-12.
47. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011;30(10):1105-17.
48. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. *BMJ* 2016;353:i3139.
49. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med* 2011;9:103.
50. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013;66(3):268-77.
51. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9(5):1-12.
52. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak* 2006;6:38.
53. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Med* 2010;8:20.

54. Debray TP, Koffijberg H, Vergouwe Y, Moons KG, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. *Stat Med* 2012;31(23):2697-712.
55. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* 2013;32(18):3158-80.
56. Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS Med* 2015;12(10):e1001886.
57. White IR, Rapsomaniki E. Covariate-adjusted measures of discrimination for survival data. *Biom J* 2015;57(4):592-613.

Supplemental material

Supplement 1: Search string

(clinical prediction[ti] OR
risk calculator*[ti] OR
risk index[ti] OR
risk indices[ti] OR
risk model*[ti] OR
risk prediction[ti] OR
risk score*[ti] OR
risk stratification[ti] OR
predictive model*[ti] OR
prediction model*[ti] OR
prediction rule*[tiab] OR
prognostic index[ti] OR
prognostic indices[ti] OR
prognostic model*[ti] OR
scoring system*[ti]) AND
(review[Publication Type] OR
review[ti] OR
critical appraisal[ti] OR
Bibliography[Publication Type] OR
Meta-analysis[Publication Type]) NOT
(Editorial[Publication Type] OR
Letter[Publication Type] OR
News[Publication Type])

Supplement 2: Description of items extracted from studies and included in analyses

Item	Extracted from studies	Categorization / handling in analyses	Description / examples
Validated model	ABCD2, ESRS, EuroSCORE, Framingham Wilson, FRAX, ISS, MELD, PSI, RCRI, SAPS 3	-	-
Study type	Predesigned validation study	Predesigned validation study	Study designed with the aim of validating a prediction model
	Validation study using existing data	Validation study using existing data	Study in which a prediction model is validated using a dataset collected for a different purpose than validating the model
	Development of new model and validation of different model	Development of new model and validation of different model	Study in which a model is developed and a model is validated
	Validation and incremental value	Validation and incremental value	Study in which a model is validated and in which the added value of one or more predictors is assessed
	Development, validation, and incremental value study	Development, validation, and incremental value study	Combination of the two above
Independent investigators	Yes	Yes	None of the authors of the development study was listed as author in the external validation study
	No	No	One or more of the authors of the development study was listed as author in the external validation study
Study design	Prospective cohort	Prospective cohort	
	Existing cohort	Existing cohort	
	Existing RCT	Existing RCT	
	Existing registry / medical records	Existing registry	
	Case-control	Case-control	
	Other (specify)	Other	

Supplement 2: Continued

Item	Extracted from studies	Categorization / handling in analyses	Description / examples
Eligibility criteria for participants	Copy/paste eligibility criteria of validation study	Comparable	Eligibility criteria comparable to development study
		Narrower	People included in the development study excluded in the validation study
		Broader	People excluded in the development stud included in the validation study
		Mixture	Combination of narrower and broader
		Unclear	
Setting	Primary care	Comparable	Same setting as development study
	Secondary care Tertiary care Population based	Broader	Same setting as development study, and participants from additional settings recruited
	Screening Mixed	Non-overlapping	Setting in development study differs from validation study
	Unclear	Unclear	
Study dates	Start year of recruitment End year of recruitment	Continuous, standardized per systematic review	
Prediction horizon	Time period for which predictions were made, e.g. 10 years.	Continuous, standardized per systematic review	
Geographical location	Country and continent	Comparable	Model validated in the same continent as the development study
		Broader	Model validated in the same and additional continents as the development study
		Non-overlapping	Model validated in a different continent than the development study
Number of centers	Number of centers (numerical)	Single	

Supplement 2: Continued

Item	Extracted from studies	Categorization / handling in analyses	Description / examples
		Multiple	
		Population based	Participants not recruited at medical centres, but, for example, from a specific geographic area (e.g. all individuals living in Framingham, US)
		Unclear	
Case-mix: age mean and sd	Mean and SD of age of participants included in the study, or other available information about age distribution	Continuous, standardized per systematic review	
Case-mix: gender	Percentage of men included in a study	Continuous, standardized per systematic review	
Predictors	Were predictors deleted from the	Yes	Changes made to predictors
	model, or were predictors	No	No changes made to predictors
	substituted with different predictors.	Unclear	
Predicted outcome	Full definition, including ICD-codes	Yes	Outcome definition comparable to development study
		No	Outcome definition not comparable to development study
		Unclear	
Outcome - measurement method	Measurement method (e.g. self-reported, interviews, expert	Yes	Outcome measurement similar for all participant
	panel), differences in outcome measurement between	No	Systematic differences in outcome measurement between participants
	participants in the study	Unclear	
Missing data	Number of participants with missing data, method of handling	Appropriate	Missing data handled using multiple imputation, or <5% missing data (arbitrary cut-off)

Supplement 2: Continued

Item	Extracted from studies	Categorization / handling in analyses	Description / examples
	missing data	Inappropriate	Missing data not handled using multiple imputation (e.g. complete-case analysis, mean imputation), and >=5% missing data
		Unclear	Unclear handling of missing data, and >=5% missing data
Number of participants		Continuous, standardized per systematic review	
Number of events		Continuous, standardized per systematic review	
Model updating	Was the model altered before validating it, e.g. using intercept recalibration.	NA	
Performance - c-statistic	C-statistic, AUC, 95% confidence intervals or SE	Logit transformation[1]	
Performance - OE ratio	OE ratio, predicted risks, presence of calibration plots or tables, 95% confidence intervals or SE	Ln transformation[1]	

SD: standard deviation, NA: not applicable, C-statistic: concordance statistic, AUC: area under the receiver operating curve, SE: standard error, OE ratio: observed expected ratio.

Information regarding c-statistics and OE ratios when not reported was sometimes restored from other information reported in the paper. If the precision of the c-statistic was not reported, we estimated this from the c-statistic and sample size of the study, using the formula described by Newcombe and Hanley.[2,3] Various equations were used to estimate the standard error of the OE ratio, depending on which information was reported. All equations (as numbered) are described in the appendix of Debray et al.[4] If the SE of the OE ratio was reported, we used equation 16 to estimate the SE of ln(OE), if the observed event risk (Po), the expected event risk (Pe), and the SE of Po were reported, we used equation 51, and if only Po and Pe were reported we used equation 27.

Supplement 3: Statistical analyses

First we pooled the total OE ratio and c-statistic within each systematic review. Based on previous recommendations,[1,4] we pooled the log OE ratio and logit c-statistic using random-effects meta-analysis accounting for the presence of between-study heterogeneity, weighted by the inverse of the variance. We calculated 95% confidence intervals (CI) and (approximate) 95% prediction intervals (PI) to quantify uncertainty and the presence of between-study heterogeneity. The Hartung-Knapp-Sidik-Jonkman (HKSJ) method was used when calculating 95% CIs.[5] The 95% PI was calculated using the equation described previously.[4] The CI indicates the precision of the summary performance estimate and the PI provides boundaries on the likely performance in future model validation studies that are comparable to the studies included in the meta-analysis, and can thus be seen as an indication of model generalizability.[6]

To study the possible association between study characteristics and predictive performance, we used a two-stepped approach. In the first stage, we fitted a univariable meta-regression model (i.e. a separate model for every study characteristic) within every systematic review, with the logit c-statistic or log OE ratio as outcome variable. This model was fitted with intercept term. Therefore, the effect estimates obtained from this meta-regression model indicate the difference in logit c-statistic or log OE ratio between a certain category of a study characteristic and a chosen reference category of that characteristic.

In the second stage, these effect estimates were pooled with a random effects meta-analysis model. This reflected the influence of the study characteristic on model performance over all systematic reviews. For continuous study characteristics, the intercept term and beta-coefficient from the first stage were jointly pooled across reviews using bivariate meta-analysis.[4,6] For categorical study characteristics the data available were not sufficient for the complexity of a multivariate model, so every category was pooled in a separate (univariate) meta-analysis.

As the estimates obtained with this approach are on the transformed scale (i.e. the difference in logit c-statistic or log OE ratio between one category and the reference category) and are difficult to be transformed back, we performed a second analysis. Here we again fitted a univariable meta-regression model, with the logit c-statistic or log OE ratio as outcome variable, but now without intercept term. This analysis enables the calculation of an effect estimate for every category of a study characteristic and to back transform this to the original scale, yielding a pooled c-statistic or pooled OE ratio for each category of a study characteristic.

We planned to perform multivariable analyses to assess the association between various study characteristics in combination and the performance of prediction models, but due to the paucity of data we were not able to do so. All analyses were performed in R version 3.3.2,[7] using the packages metafor,[8] mvmeta,[9] metamisc,[10] and lme4.[11]

Table S1: Description of study characteristics and quality of reporting within each systematic review

Categorical variables

	ABCD2	ESRS	EuroSCORE	Framingham	FRAX	ISS	MELD	PSI	RCRI	SAPS 3
Studytype										
Validation study using existing data	9 (56%)	7 (64%)	21 (95%)	18 (78%)	26 (87%)	24 (71%)	10 (71%)	16 (67%)	11 (48%)	8 (30%)
Development of new model and validation of different model	2 (12%)	2 (18%)	1 (5%)	2 (9%)	1 (3%)	4 (12%)	3 (21%)	2 (8%)	2 (9%)	6 (22%)
Development, validation, and incremental value study	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (6%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Predesigned validation study	1 (6%)	1 (9%)	0 (0%)	0 (0%)	1 (3%)	1 (3%)	0 (0%)	5 (21%)	1 (4%)	10 (37%)
Validation and incremental value	4 (25%)	1 (9%)	0 (0%)	3 (13%)	2 (7%)	3 (9%)	1 (7%)	1 (4%)	9 (39%)	3 (11%)
Study design										
Existing cohort	12 (75%)	4 (36%)	3 (14%)	20 (87%)	20 (67%)	1 (3%)	2 (14%)	4 (17%)	7 (30%)	1 (4%)
Prospective cohort	1 (6%)	2 (18%)	0 (0%)	0 (0%)	1 (3%)	3 (9%)	0 (0%)	6 (25%)	4 (17%)	17 (63%)
Existing RCT	0 (0%)	1 (9%)	0 (0%)	0 (0%)	3 (10%)	0 (0%)	0 (0%)	0 (0%)	2 (9%)	0 (0%)
Existing registry	3 (19%)	4 (36%)	19 (86%)	1 (4%)	2 (7%)	30 (88%)	12 (86%)	14 (58%)	10 (43%)	9 (33%)
Case-control	0 (0%)	0 (0%)	0 (0%)	0 (0%)	4 (13%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Other	0 (0%)	0 (0%)	0 (0%)	2 (9%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Validation by independent investigators										
No	6 (38%)	3 (27%)	3 (14%)	10 (43%)	17 (57%)	2 (6%)	0 (0%)	7 (29%)	1 (4%)	2 (7%)

Table S1: Continued

	ABCD2	ESRS	EuroSCORE	Framingham	FRAX	ISS	MELD	PSI	RCRI	SAPS 3
Yes	10 (62%)	8 (73%)	19 (86%)	13 (57%)	13 (43%)	32 (94%)	14 (100%)	17 (71%)	22 (96%)	25 (93%)
Comparability of eligibility criteria										
Narrower	6 (38%)	2 (18%)	18 (82%)	18 (78%)	28 (93%)	22 (65%)	0 (0%)	4 (17%)	4 (17%)	20 (74%)
Comparable	4 (25%)	0 (0%)	2 (9%)	3 (13%)	2 (7%)	5 (15%)	0 (0%)	1 (4%)	3 (13%)	7 (26%)
Mixture	5 (31%)	9 (82%)	0 (0%)	2 (9%)	0 (0%)	3 (9%)	0 (0%)	16 (67%)	11 (48%)	0 (0%)
Broader	1 (6%)	0 (0%)	2 (9%)	0 (0%)	0 (0%)	2 (6%)	0 (0%)	3 (12%)	5 (22%)	0 (0%)
Non-overlapping	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	14 (100%)	0 (0%)	0 (0%)	0 (0%)
Unclear	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (6%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Comparability of age eligibility criteria										
Narrower	1 (6%)	2 (18%)	0 (0%)	17 (74%)	9 (30%)	13 (38%)	10 (71%)	2 (8%)	1 (4%)	17 (63%)
Comparable	15 (94%)	0 (0%)	22 (100%)	0 (0%)	1 (3%)	21 (62%)	4 (29%)	15 (62%)	4 (17%)	4 (15%)
Mixture	0 (0%)	0 (0%)	0 (0%)	6 (26%)	16 (53%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Broader	0 (0%)	9 (82%)	0 (0%)	0 (0%)	4 (13%)	0 (0%)	0 (0%)	7 (29%)	18 (78%)	6 (22%)
Setting										
Primary care	3 (19%)	0 (0%)	0 (0%)	7 (30%)	3 (10%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Secondary care	12 (75%)	9 (82%)	16 (84%)	0 (0%)	5 (17%)	18 (82%)	6 (75%)	18 (90%)	12 (86%)	4 (40%)
Tertiary care	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Population based	0 (0%)	1 (9%)	0 (0%)	15 (65%)	17 (59%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Screening	0 (0%)	0 (0%)	0 (0%)	1 (4%)	1 (3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Table S1: Continued

	ABCD2	ESRS	EuroSCORE	Framingham	FRAX	ISS	MELD	PSI	RCRI	SAPS 3
Mixed	1 (6%)	1 (9%)	2 (11%)	0 (0%)	0 (0%)	4 (18%)	1 (12%)	2 (10%)	2 (14%)	2 (20%)
Unclear	0 (0%)	0 (0%)	1 (5%)	0 (0%)	3 (10%)	0 (0%)	1 (12%)	0 (0%)	0 (0%)	4 (40%)
Comparability of setting										
Comparable	1 (6%)	0 (0%)	16 (73%)	15 (65%)	17 (57%)	18 (53%)	6 (43%)	18 (75%)	9 (39%)	4 (15%)
Narrower	15 (94%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Broader	0 (0%)	0 (0%)	2 (9%)	0 (0%)	0 (0%)	4 (12%)	1 (7%)	2 (8%)	2 (9%)	2 (7%)
Non-overlapping	0 (0%)	11 (100%)	3 (14%)	8 (35%)	10 (33%)	12 (35%)	6 (43%)	4 (17%)	12 (52%)	17 (63%)
Unclear	0 (0%)	0 (0%)	1 (5%)	0 (0%)	3 (10%)	0 (0%)	1 (7%)	0 (0%)	0 (0%)	4 (15%)
Continent										
Africa	0 (0%)	0 (0%)	1 (5%)	0 (0%)	0 (0%)	1 (3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Asia	0 (0%)	2 (18%)	4 (18%)	2 (9%)	4 (13%)	7 (21%)	1 (7%)	3 (12%)	2 (9%)	5 (19%)
Australia	0 (0%)	0 (0%)	1 (5%)	0 (0%)	5 (17%)	2 (6%)	0 (0%)	3 (12%)	1 (4%)	1 (4%)
Europe	8 (50%)	7 (64%)	10 (45%)	10 (43%)	9 (30%)	7 (21%)	7 (50%)	11 (46%)	13 (57%)	9 (33%)
North America	8 (50%)	1 (9%)	5 (23%)	11 (48%)	11 (37%)	17 (50%)	3 (21%)	6 (25%)	6 (26%)	3 (11%)
South America	0 (0%)	0 (0%)	1 (5%)	0 (0%)	0 (0%)	0 (0%)	3 (21%)	0 (0%)	0 (0%)	9 (33%)
Combination	0 (0%)	1 (9%)	0 (0%)	0 (0%)	1 (3%)	0 (0%)	0 (0%)	1 (4%)	1 (4%)	0 (0%)
Comparability of continent										
Comparable	0 (0%)	0 (0%)	10 (45%)	11 (48%)	0 (0%)	17 (50%)	3 (21%)	6 (25%)	6 (26%)	0 (0%)

Table S1: Continued

	ABCD2	ESRS	EuroSCORE	Framingham	FRAX	ISS	MELD	PSI	RCRI	SAPS 3
Narrower	16 (100%)	8 (73%)	0 (0%)	0 (0%)	30 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	27 (100%)
Broader	0 (0%)	1 (9%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (4%)	1 (4%)	0 (0%)
Non-overlapping	0 (0%)	2 (18%)	12 (55%)	12 (52%)	0 (0%)	17 (50%)	11 (79%)	17 (71%)	16 (70%)	0 (0%)
Number of centers										
Single	9 (56%)	4 (36%)	12 (55%)	3 (13%)	5 (17%)	18 (53%)	12 (86%)	14 (58%)	17 (74%)	13 (48%)
Multiple	6 (38%)	7 (64%)	9 (41%)	6 (26%)	9 (30%)	15 (44%)	2 (14%)	10 (42%)	6 (26%)	14 (52%)
Population based	1 (6%)	0 (0%)	0 (0%)	12 (52%)	15 (50%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Unclear	0 (0%)	0 (0%)	1 (5%)	2 (9%)	1 (3%)	1 (3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Changes made to predictors										
No	16 (100%)	10 (91%)	13 (59%)	23 (100%)	20 (67%)	21 (62%)	12 (86%)	13 (54%)	17 (74%)	26 (96%)
Yes	0 (0%)	0 (0%)	5 (23%)	0 (0%)	10 (33%)	0 (0%)	0 (0%)	5 (21%)	2 (9%)	0 (0%)
Unclear	0 (0%)	1 (9%)	4 (18%)	0 (0%)	0 (0%)	13 (38%)	2 (14%)	6 (25%)	4 (17%)	1 (4%)
Comparability of outcome definition										
No	8 (50%)	3 (27%)	9 (41%)	13 (57%)	16 (53%)	7 (21%)	14 (100%)	5 (21%)	5 (22%)	24 (89%)
Yes	3 (19%)	8 (73%)	11 (50%)	4 (17%)	13 (43%)	19 (56%)	0 (0%)	18 (75%)	18 (78%)	3 (11%)
Unclear	5 (31%)	0 (0%)	2 (9%)	6 (26%)	1 (3%)	8 (24%)	0 (0%)	1 (4%)	0 (0%)	0 (0%)
Outcome measurement method										
Self-reported	3 (19%)	8 (73%)	2 (9%)	15 (65%)	18 (60%)	0 (0%)	1 (7%)	6 (25%)	1 (4%)	2 (7%)
Clinician	6 (38%)	1 (9%)	2 (9%)	2 (9%)	1 (3%)	4 (12%)	0 (0%)	2 (8%)	10 (43%)	6 (22%)

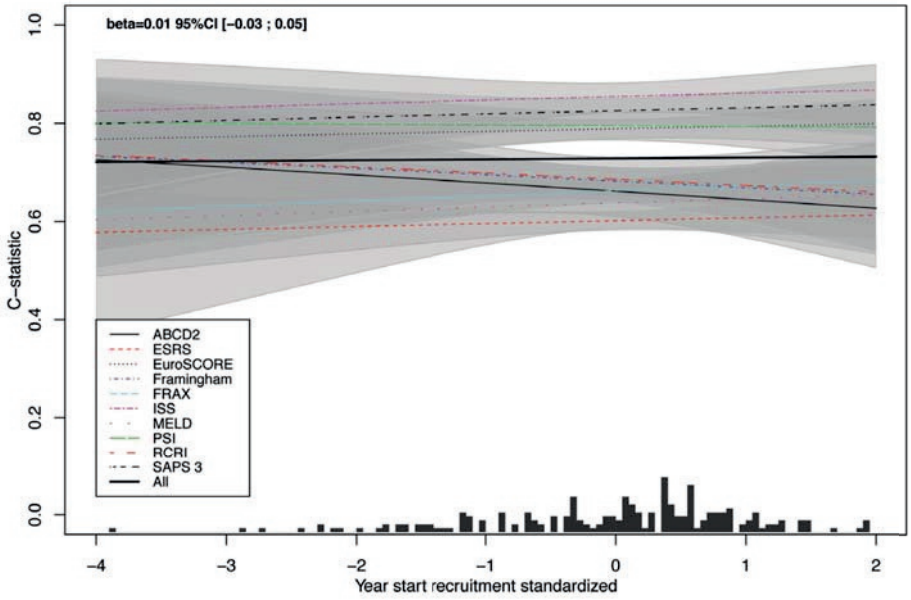
Table S1: Continued

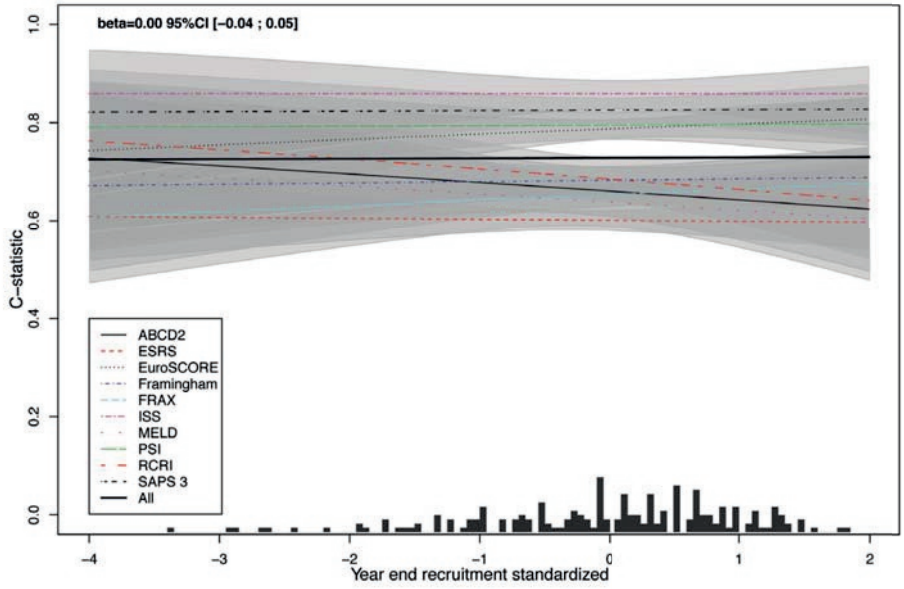
	ABCD2	ESRS	EuroSCORE	Framingham	FRAX	ISS	MELD	PSI	RCRI	SAPS 3
Registry	3 (19%)	2 (18%)	5 (23%)	4 (17%)	8 (27%)	13 (38%)	3 (21%)	9 (38%)	6 (26%)	9 (33%)
Unclear	4 (25%)	0 (0%)	13 (59%)	2 (9%)	3 (10%)	17 (50%)	10 (71%)	7 (29%)	6 (26%)	10 (37%)
Similar outcome measurement for all patients										
Yes	9 (56%)	3 (27%)	11 (50%)	6 (26%)	14 (47%)	18 (53%)	2 (14%)	12 (50%)	13 (57%)	16 (59%)
No	1 (6%)	5 (45%)	1 (5%)	16 (70%)	3 (10%)	3 (9%)	0 (0%)	3 (12%)	4 (17%)	1 (4%)
Unclear	6 (38%)	3 (27%)	10 (45%)	1 (4%)	13 (43%)	13 (38%)	12 (86%)	9 (38%)	6 (26%)	10 (37%)
Method for handling of missing data										
Complete case analysis	4 (25%)	8 (73%)	3 (14%)	11 (48%)	19 (63%)	18 (53%)	7 (50%)	1 (4%)	5 (22%)	8 (30%)
Mean/median imputation	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (4%)
Multiple imputation	1 (6%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
NA	3 (19%)	1 (9%)	1 (5%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (4%)	1 (4%)	2 (7%)
Other	0 (0%)	0 (0%)	1 (5%)	0 (0%)	2 (7%)	0 (0%)	1 (7%)	8 (33%)	0 (0%)	7 (26%)
Unclear	8 (50%)	2 (18%)	17 (77%)	12 (52%)	9 (30%)	16 (47%)	6 (43%)	14 (58%)	17 (74%)	9 (33%)
Handling of missing data										
Appropriate or <5% missing	8 (50%)	4 (36%)	2 (9%)	3 (13%)	4 (13%)	6 (18%)	4 (29%)	4 (17%)	7 (30%)	6 (22%)
Inappropriate	1 (6%)	5 (45%)	3 (14%)	8 (35%)	17 (57%)	13 (38%)	4 (29%)	8 (33%)	0 (0%)	12 (44%)
Unclear	7 (44%)	2 (18%)	17 (77%)	12 (52%)	9 (30%)	15 (44%)	6 (43%)	12 (50%)	16 (70%)	9 (33%)

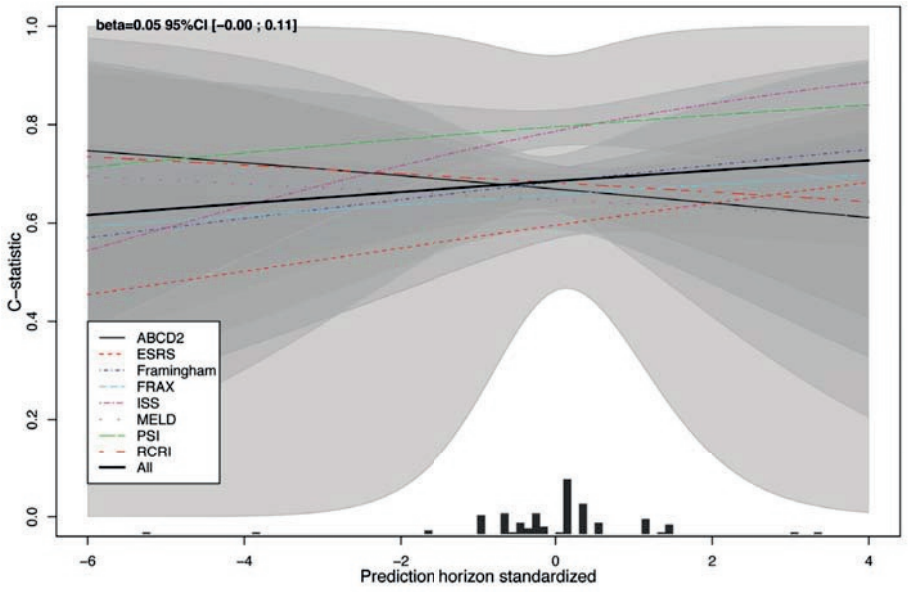
Continuous variables

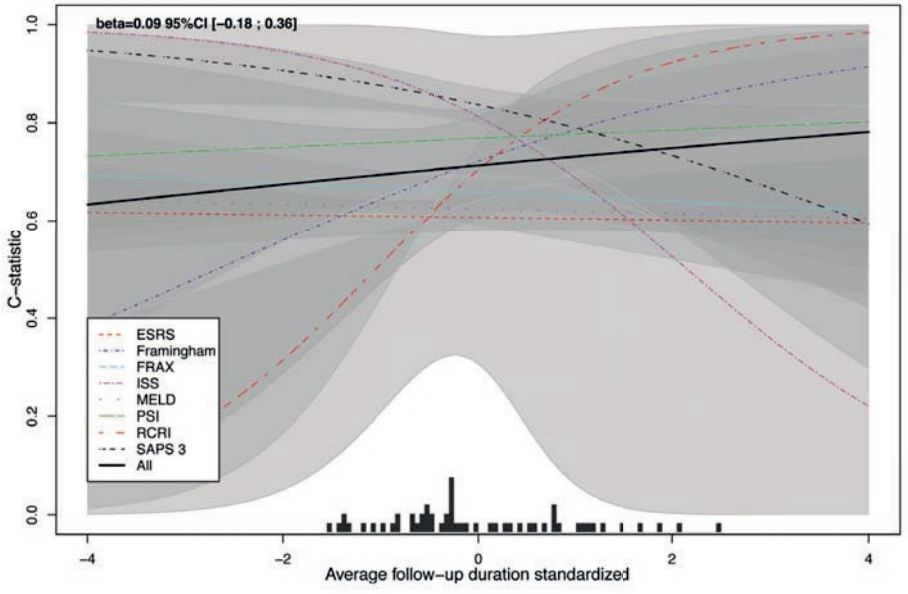
	ABCD2	ESRS	EuroSCORE	Framingham	FRAX	ISS	MELD	PSI	RCRI	SAPSIII
Year start recruitment	2002 (2000-2003) NR=0	2007 (2004-2007) NR=0	1998 (1995-2001) NR=1	1989 (1983-1994) NR=0	1994 (1990-1998) NR=3	1996 (1993-1998) NR=1	2000 (1998-2004) NR=0	2000 (1998-2002) NR=0	2000 (1994-2002) NR=4	2006 (2006-2007) NR=0
Year end recruitment	2005 (2003-2007) NR=0	2008 (2006-2008) NR=0	2002 (1999-2005) NR=1	1993 (1988-1998) NR=0	1997 (1993-2006) NR=8	2000 (1996-2003) NR=2	2006 (2004-2007) NR=0	2002 (2000-2003) NR=0	2002 (2000-2005) NR=4	2007 (2006-2009) NR=0
Percentage missings	0.95 (0.00-5.00) NR=7	5.12 (1.99-17.80) NR=2	6.40 (1.50-11.83) NR=18	4.90 (2.70-9.80) NR=18	30.25 (2.75-33.80) NR=16	9.05 (2.40-14.65) NR=20	4.05 (2.73-10.93) NR=8	0.52 (0.07-9.26) NR=18	1.00 (0.09-1.91) NR=16	5.85 (0.52-18.93) NR=15
Number of participants	304 (204-691) NR=0	1257 (712-2594) NR=0	1730 (873-4518) NR=2	2399 (928-4609) NR=0	2210 (889-6586) NR=0	2590 (960-20713) NR=0	418 (118-483) NR=0	730 (326-970) NR=1	496 (180-1480) NR=0	864 (485-1856) NR=0
Number of events	9 (3-18) NR=0	92 (60-134) NR=0	36 (13-87) NR=2	92 (72-160) NR=1	250 (86-581) NR=0	256 (113-1660) NR=2	49 (22-112) NR=0	54 (28-111) NR=1	31 (14-76) NR=0	180 (124-311) NR=1
Age mean	67.4 (64.1-70.0) NR=5	68.3 (67.1-71.5) NR=3	63.9 (62.5-65.2) NR=2	54.6 (50.9-58.3) NR=2	66.8 (63.0-71.3) NR=1	38.1 (32.4-41.3) NR=10	51.8 (49.1-53.0) NR=0	66.2 (64.0-69.3) NR=2	67.8 (66.0-71.9) NR=2	62.2 (60.8-64.8) NR=1
Age sd	13.8 (13.0-14.9) NR=5	12.4 (12.0-13.0) NR=1	9.3 (9.0-10.6) NR=8	7.3 (4.1-9.4) NR=0	8.3 (5.9-9.8) NR=0	20.9 (18.1-24.8) NR=2	10.0 (9.6-12.0) NR=1	17.8 (17.0-20.1) NR=3	10.0 (8.8-12.5) NR=4	17.0 (15.4-19.0) NR=3
Gender percentage men	47 (45-53) NR=2	57 (55-59) NR=1	77 (71-79) NR=1	100 (100-100) NR=0	0 (0-0) NR=0	71 (64-75) NR=11	68 (63-69) NR=0	57 (53-64) NR=1	67 (52-76) NR=0	59 (55-64) NR=0

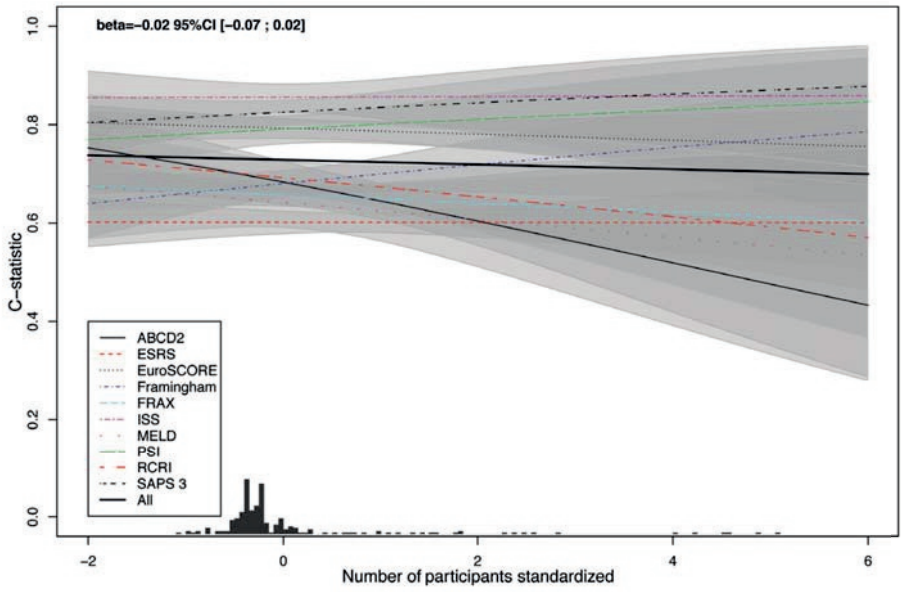
Figure S1: Associations between continuous variables and c-statistic

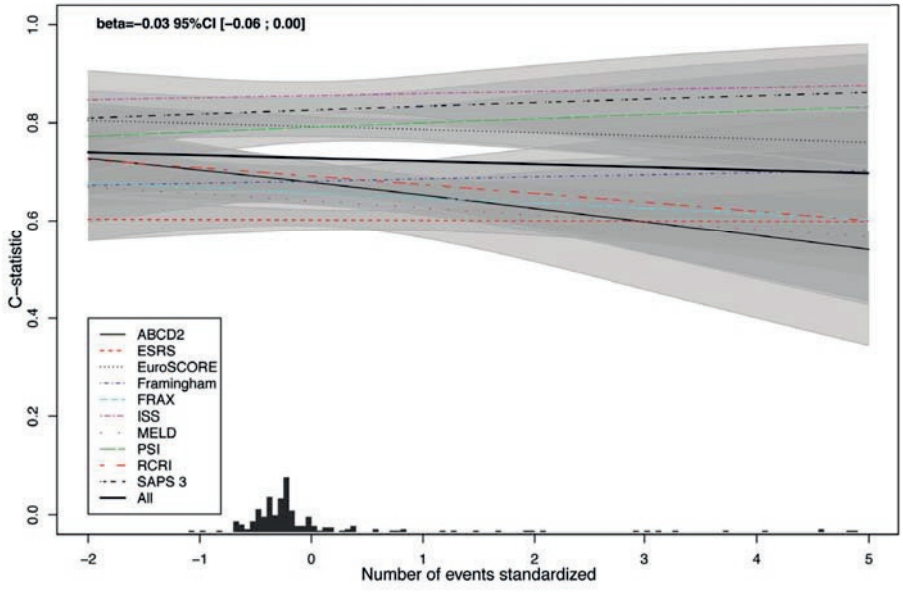


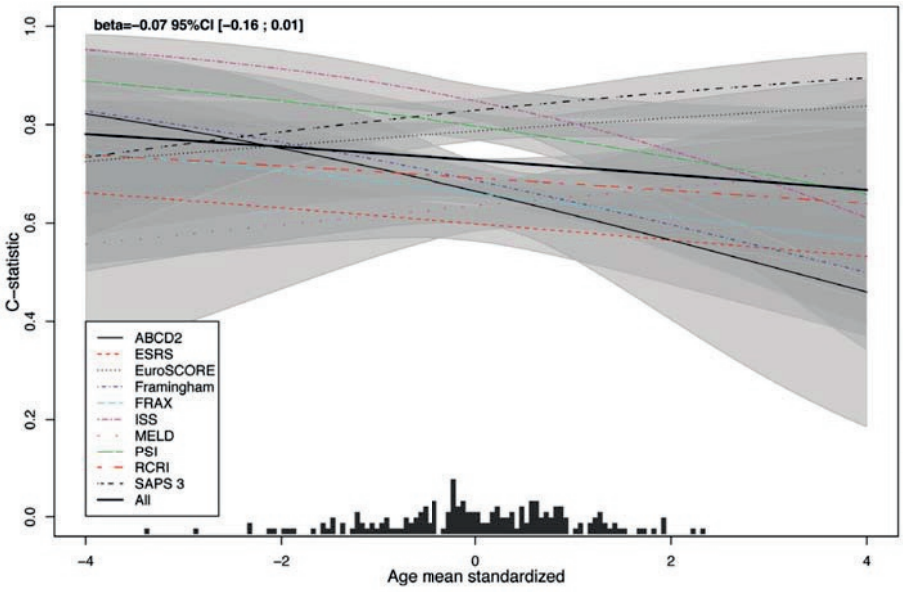


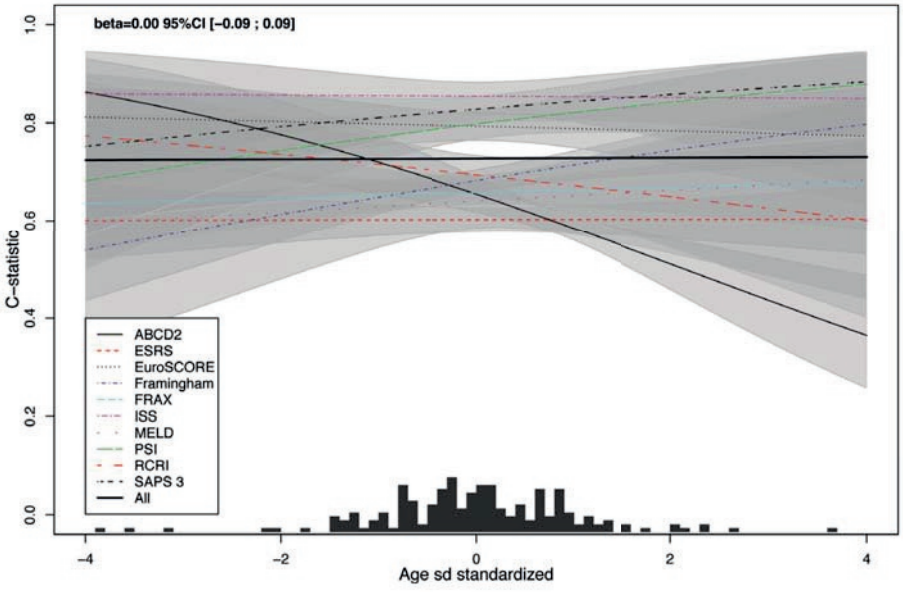

beta=0.05 95%CI [−0.00 ; 0.11]
C−statistic
Prediction horizon standardized
ABCD2
ESRS
Framingham
FRAX
ISS
MELD
PSI
RCRI
All


beta=0.09 95%CI [−0.18 ; 0.36]
C−statistic
Average follow−up duration standardized
ESRS
Framingham
FRAX
ISS
MELD
PSI
RCRI
SAPS 3
All


beta=−0.02 95%CI [−0.07 ; 0.02]
C−statistic
1.0
0.8
0.6
0.4
0.2
0.0
ABCD2
ESRS
EuroSCORE
Framingham
FRAX
ISS
MELD
PSI
RCRI
SAPS 3
All
−2
0
2
4
6
Number of participants standardized


beta=−0.03 95%CI [−0.06 ; 0.00]
C−statistic
1.0
0.8
0.6
0.4
0.2
0.0
ABCD2
ESRS
EuroSCORE
Framingham
FRAX
ISS
MELD
PSI
RCRI
SAPS 3
All
−2
−1
0
1
2
3
4
5
Number of events standardized


beta=−0.07 95%CI [−0.16 ; 0.01]
C−statistic
Age mean standardized
ABCD2
ESRS
EuroSCORE
Framingham
FRAX
ISS
MELD
PSI
RCRI
SAPS 3
All


beta=0.00 95%CI [−0.09 ; 0.09]
C−statistic
Age sd standardized
ABCD2
ESRS
EuroSCORE
Framingham
FRAX
ISS
MELD
PSI
RCRI
SAPS 3
All

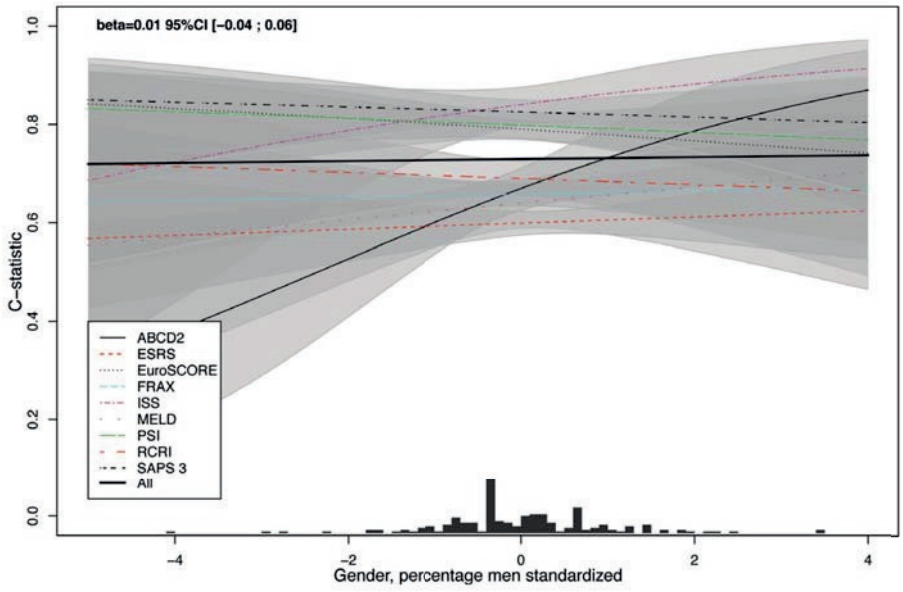
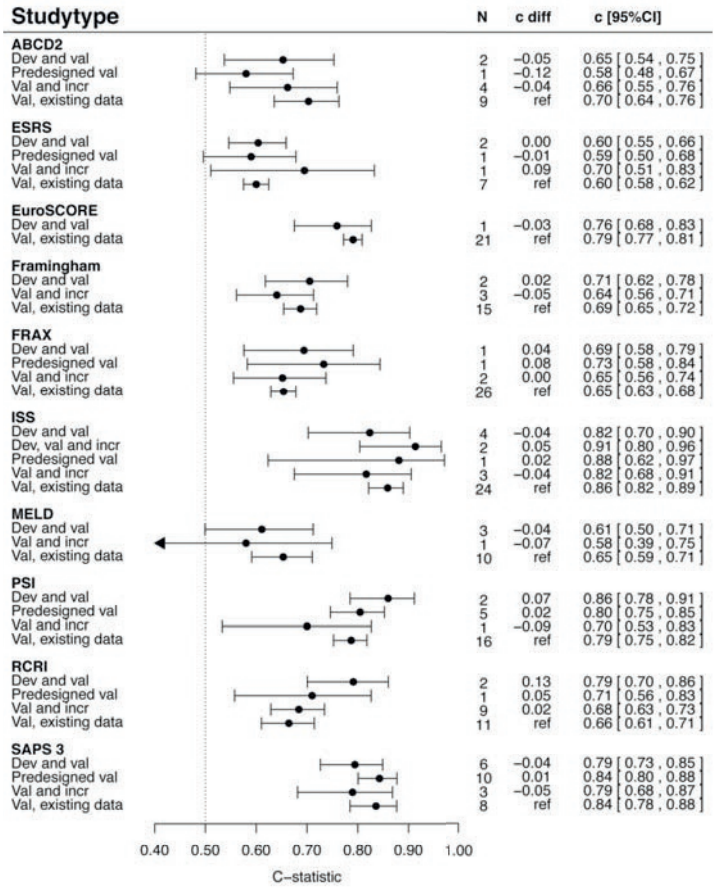
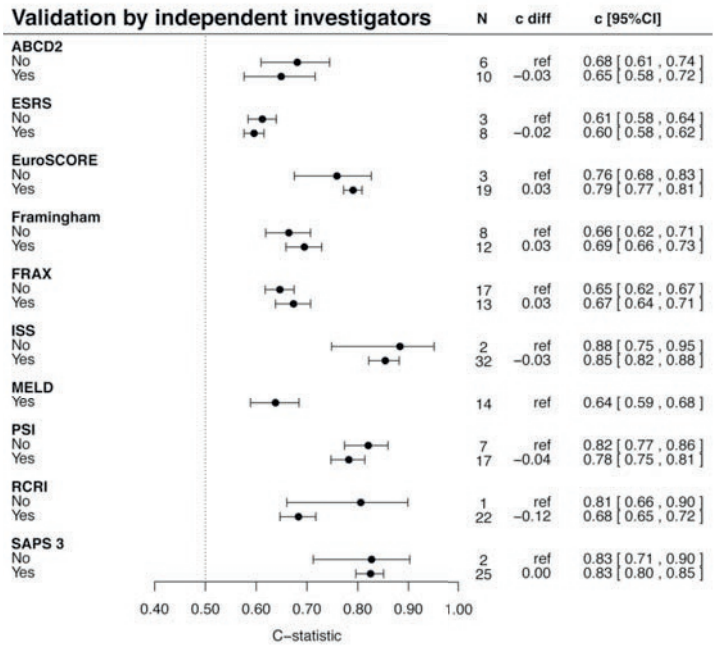
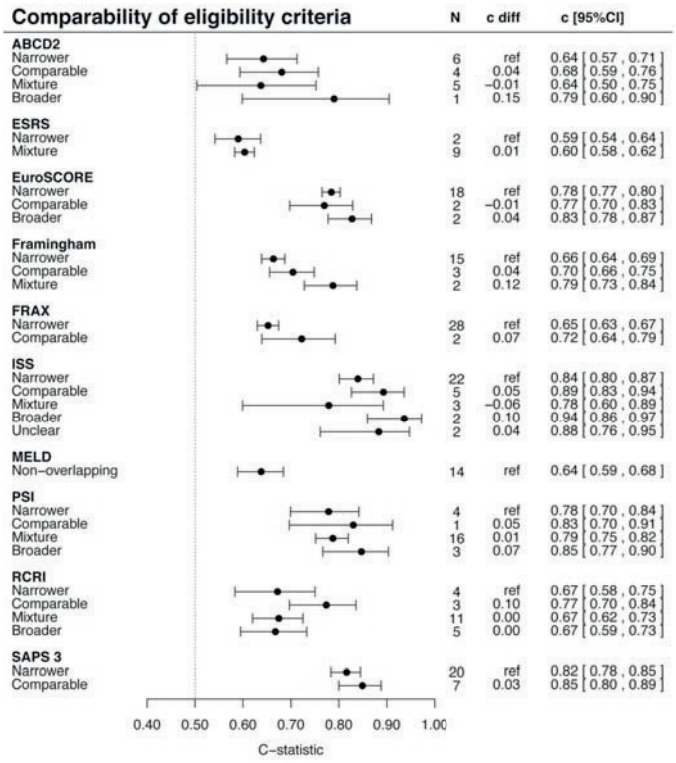
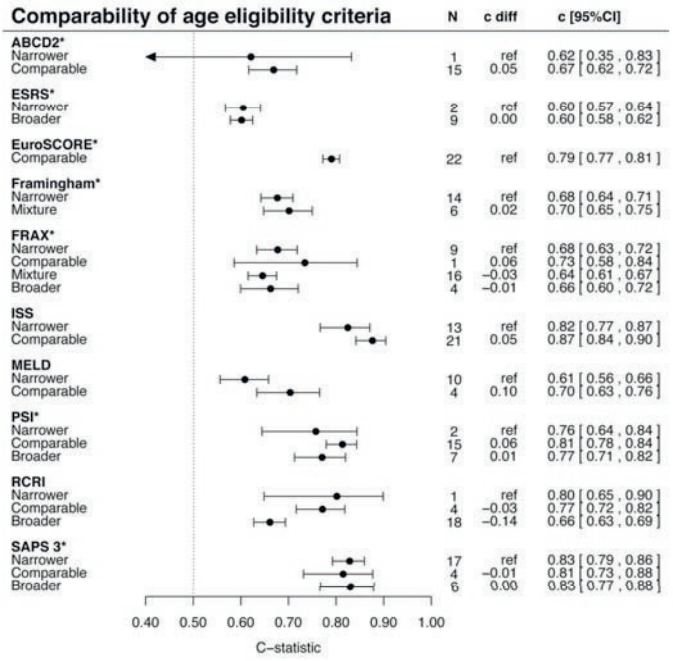

beta=0.01 95%CI [−0.04 ; 0.06]
C−statistic
0.0
0.2
0.4
0.6
0.8
1.0
ABCD2
ESRS
EuroSCORE
FRAX
ISS
MELD
PSI
RCRI
SAPS 3
All
−4
−2
0
2
4
Gender, percentage men standardized

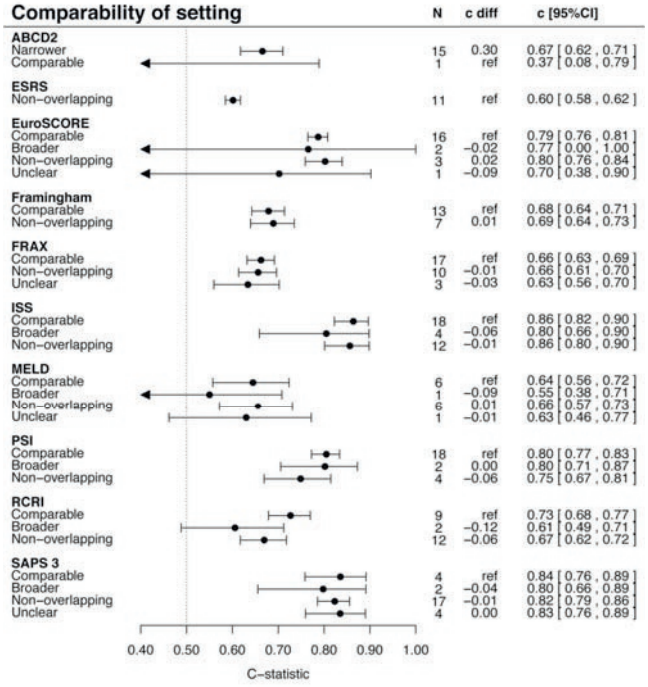
Figure S2: C-statistic in categories of study characteristics within each systematic review

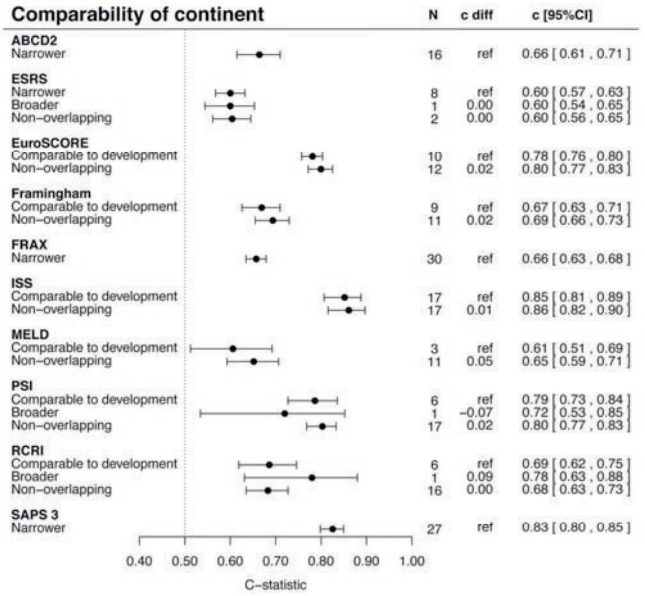


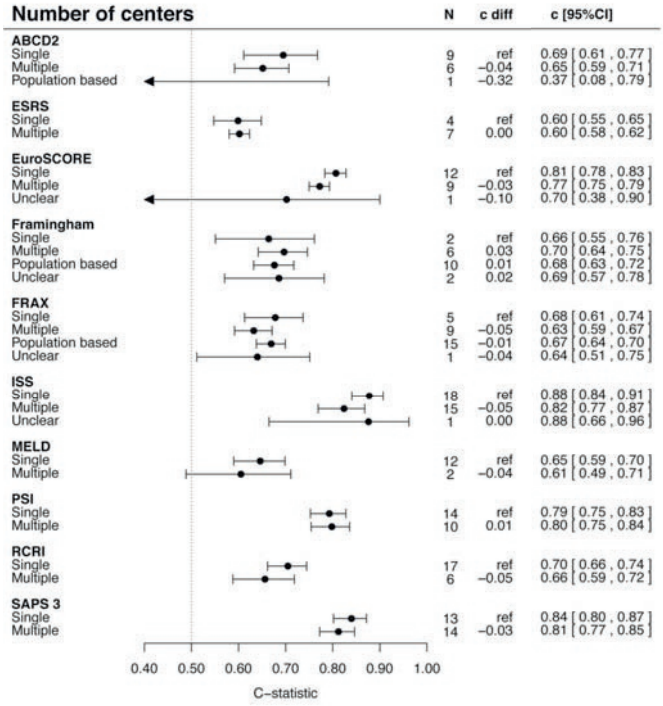


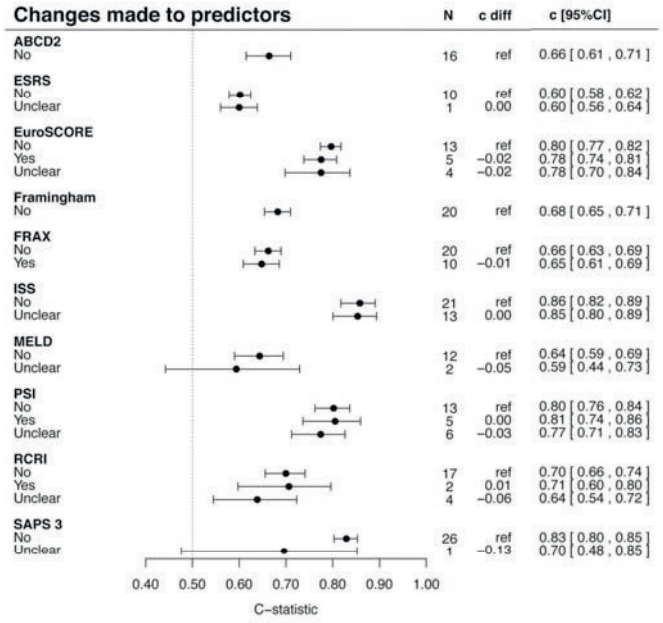

Comparability of eligibility criteria
N
c diff
c [95%CI]
ABCD2
Narrower 6 ref 0.64 [0.57 , 0.71]
Comparable 4 0.04 0.68 [0.59 , 0.76]
Mixture 5 −0.01 0.64 [0.50 , 0.75]
Broader 1 0.15 0.79 [0.60 , 0.90]
ESRS
Narrower 2 ref 0.59 [0.54 , 0.64]
Mixture 9 0.01 0.60 [0.58 , 0.62]
EuroSCORE
Narrower 18 ref 0.78 [0.77 , 0.80]
Comparable 2 −0.01 0.77 [0.70 , 0.83]
Broader 2 0.04 0.83 [0.78 , 0.87]
Framingham
Narrower 15 ref 0.66 [0.64 , 0.69]
Comparable 3 0.04 0.70 [0.66 , 0.75]
Mixture 2 0.12 0.79 [0.73 , 0.84]
FRAX
Narrower 28 ref 0.65 [0.63 , 0.67]
Comparable 2 0.07 0.72 [0.64 , 0.79]
ISS
Narrower 22 ref 0.84 [0.80 , 0.87]
Comparable 5 0.05 0.89 [0.83 , 0.94]
Mixture 3 −0.06 0.78 [0.60 , 0.89]
Broader 2 0.10 0.94 [0.86 , 0.97]
Unclear 2 0.04 0.88 [0.76 , 0.95]
MELD
Non−overlapping 14 ref 0.64 [0.59 , 0.68]
PSI
Narrower 4 ref 0.78 [0.70 , 0.84]
Comparable 1 0.05 0.83 [0.70 , 0.91]
Mixture 16 0.01 0.79 [0.75 , 0.82]
Broader 3 0.07 0.85 [0.77 , 0.90]
RCRI
Narrower 4 ref 0.67 [0.58 , 0.75]
Comparable 3 0.10 0.77 [0.70 , 0.84]
Mixture 11 0.00 0.67 [0.62 , 0.73]
Broader 5 0.00 0.67 [0.59 , 0.73]
SAPS 3
Narrower 20 ref 0.82 [0.78 , 0.85]
Comparable 7 0.03 0.85 [0.80 , 0.89]
0.40 0.50 0.60 0.70 0.80 0.90 1.00
C−statistic

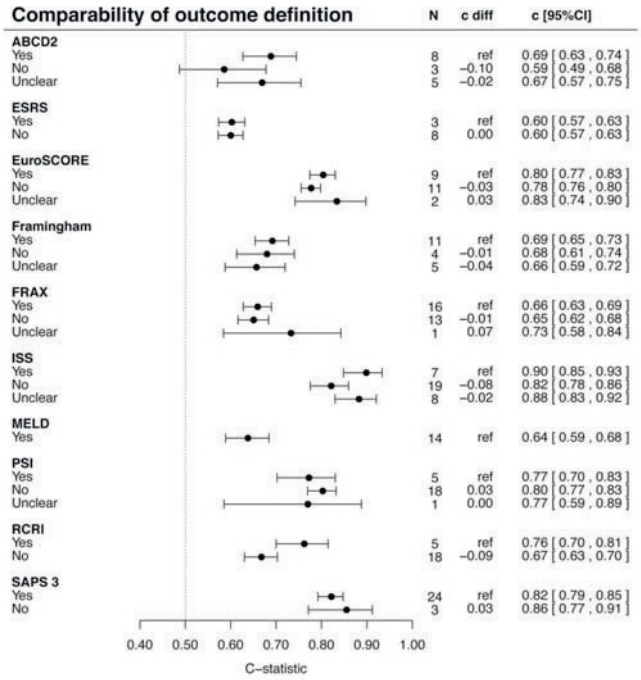

Comparability of age eligibility criteria
N
c diff
c [95%CI]
ABCD2*
Narrower 1 ref 0.62 [0.35 , 0.83]
Comparable 15 0.05 0.67 [0.62 , 0.72]
ESRS*
Narrower 2 ref 0.60 [0.57 , 0.64]
Broader 9 0.00 0.60 [0.58 , 0.62]
EuroSCORE*
Comparable 22 ref 0.79 [0.77 , 0.81]
Framingham*
Narrower 14 ref 0.68 [0.64 , 0.71]
Mixture 6 0.02 0.70 [0.65 , 0.75]
FRAX*
Narrower 9 ref 0.68 [0.63 , 0.72]
Comparable 1 0.06 0.73 [0.58 , 0.84]
Mixture 16 −0.03 0.64 [0.61 , 0.67]
Broader 4 −0.01 0.66 [0.60 , 0.72]
ISS
Narrower 13 ref 0.82 [0.77 , 0.87]
Comparable 21 0.05 0.87 [0.84 , 0.90]
MELD
Narrower 10 ref 0.61 [0.56 , 0.66]
Comparable 4 0.10 0.70 [0.63 , 0.76]
PSI*
Narrower 2 ref 0.76 [0.64 , 0.84]
Comparable 15 0.06 0.81 [0.78 , 0.84]
Broader 7 0.01 0.77 [0.71 , 0.82]
RCRI
Narrower 1 ref 0.80 [0.65 , 0.90]
Comparable 4 −0.03 0.77 [0.72 , 0.82]
Broader 18 −0.14 0.66 [0.63 , 0.69]
SAPS 3*
Narrower 17 ref 0.83 [0.79 , 0.86]
Comparable 4 −0.01 0.81 [0.73 , 0.88]
Broader 6 0.00 0.83 [0.77 , 0.88]
0.40 0.50 0.60 0.70 0.80 0.90 1.00
C−statistic

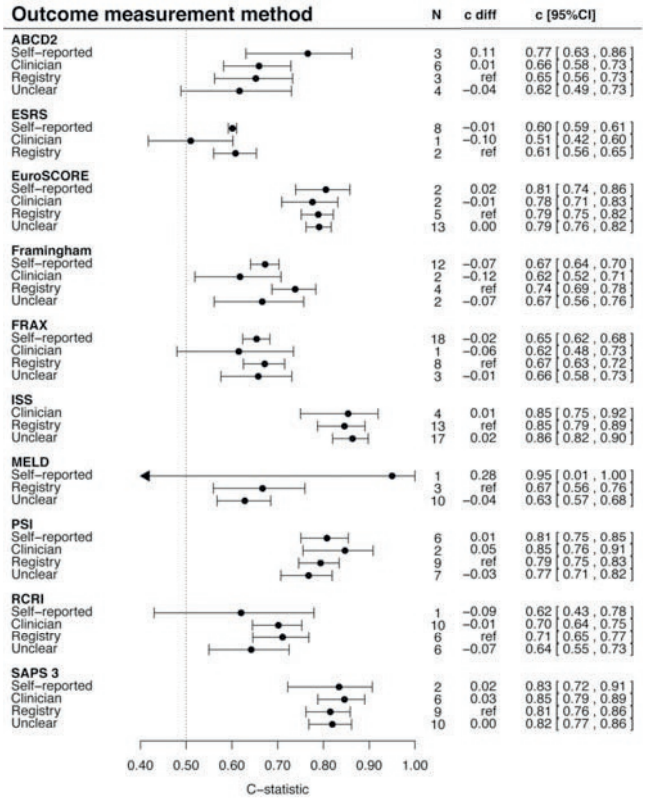

Comparability of setting
N
c diff
c [95%CI]
ABCD2
Narrower 15 0.30 0.67 [0.62 , 0.71]
Comparable 1 ref 0.37 [0.08 , 0.79]
ESRS
Non−overlapping 11 ref 0.60 [0.58 , 0.62]
EuroSCORE
Comparable 16 ref 0.79 [0.76 , 0.81]
Broader 2 −0.02 0.77 [0.00 , 1.00]
Non−overlapping 3 0.02 0.80 [0.76 , 0.84]
Unclear 1 −0.09 0.70 [0.38 , 0.90]
Framingham
Comparable 13 ref 0.68 [0.64 , 0.71]
Non−overlapping 7 0.01 0.69 [0.64 , 0.73]
FRAX
Comparable 17 ref 0.66 [0.63 , 0.69]
Non−overlapping 10 −0.01 0.66 [0.61 , 0.70]
Unclear 3 −0.03 0.63 [0.56 , 0.70]
ISS
Comparable 18 ref 0.86 [0.82 , 0.90]
Broader 4 −0.06 0.80 [0.66 , 0.90]
Non−overlapping 12 −0.01 0.86 [0.80 , 0.90]
MELD
Comparable 6 ref 0.64 [0.56 , 0.72]
Broader 1 −0.09 0.55 [0.38 , 0.71]
Non−overlapping 6 0.01 0.66 [0.57 , 0.73]
Unclear 1 −0.01 0.63 [0.46 , 0.77]
PSI
Comparable 18 ref 0.80 [0.77 , 0.83]
Broader 2 0.00 0.80 [0.71 , 0.87]
Non−overlapping 4 −0.06 0.75 [0.67 , 0.81]
RCRI
Comparable 9 ref 0.73 [0.68 , 0.77]
Broader 2 −0.12 0.61 [0.49 , 0.71]
Non−overlapping 12 −0.06 0.67 [0.62 , 0.72]
SAPS 3
Comparable 4 ref 0.84 [0.76 , 0.89]
Broader 2 −0.04 0.80 [0.66 , 0.89]
Non−overlapping 17 −0.01 0.82 [0.79 , 0.86]
Unclear 4 0.00 0.83 [0.76 , 0.89]
0.40 0.50 0.60 0.70 0.80 0.90 1.00
C−statistic

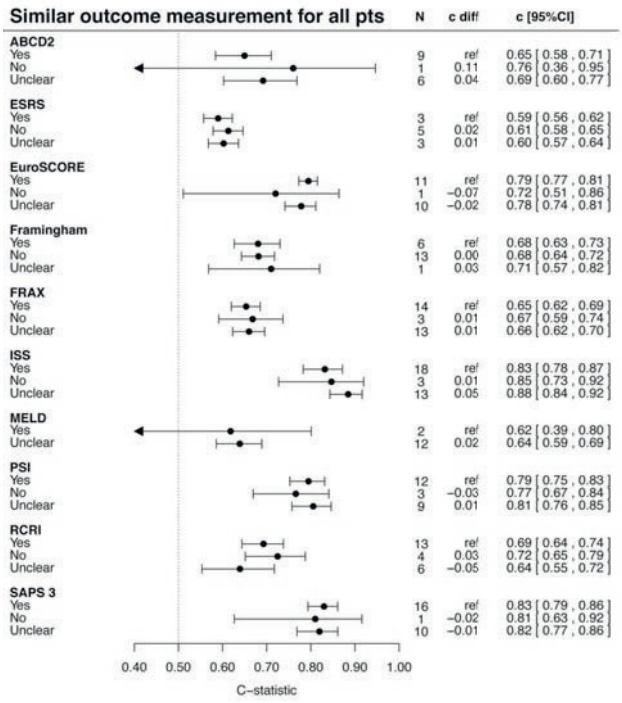

Comparability of continent
N
c diff
c [95%CI]
ABCD2
Narrower 16 ref 0.66 [0.61 , 0.71]
ESRS
Narrower 8 ref 0.60 [0.57 , 0.63]
Broader 1 0.00 0.60 [0.54 , 0.65]
Non−overlapping 2 0.00 0.60 [0.56 , 0.65]
EuroSCORE
Comparable to development 10 ref 0.78 [0.76 , 0.80]
Non−overlapping 12 0.02 0.80 [0.77 , 0.83]
Framingham
Comparable to development 9 ref 0.67 [0.63 , 0.71]
Non−overlapping 11 0.02 0.69 [0.66 , 0.73]
FRAX
Narrower 30 ref 0.66 [0.63 , 0.68]
ISS
Comparable to development 17 ref 0.85 [0.81 , 0.89]
Non−overlapping 17 0.01 0.86 [0.82 , 0.90]
MELD
Comparable to development 3 ref 0.61 [0.51 , 0.69]
Non−overlapping 11 0.05 0.65 [0.59 , 0.71]
PSI
Comparable to development 6 ref 0.79 [0.73 , 0.84]
Broader 1 −0.07 0.72 [0.53 , 0.85]
Non−overlapping 17 0.02 0.80 [0.77 , 0.83]
RCRI
Comparable to development 6 ref 0.69 [0.62 , 0.75]
Broader 1 0.09 0.78 [0.63 , 0.88]
Non−overlapping 16 0.00 0.68 [0.63 , 0.73]
SAPS 3
Narrower 27 ref 0.83 [0.80 , 0.85]
0.40 0.50 0.60 0.70 0.80 0.90 1.00
C−statistic

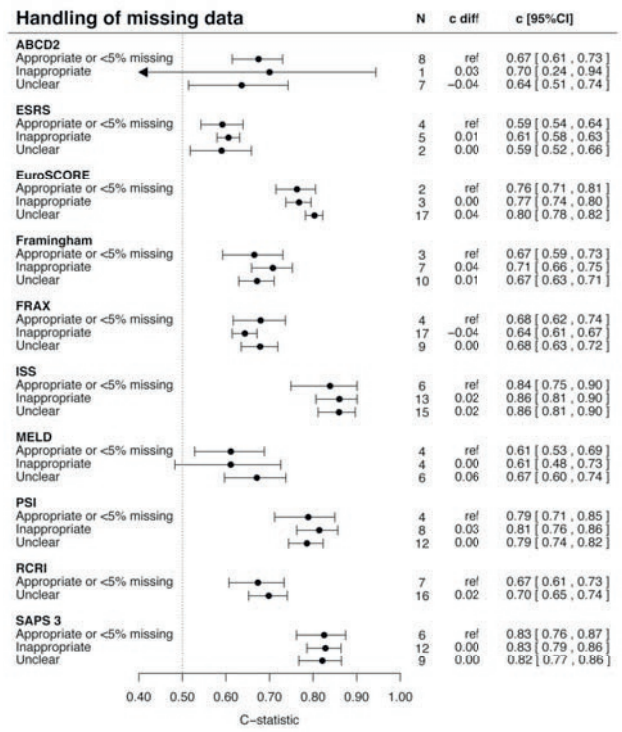

Number of centers
N
c diff
c [95%CI]
ABCD2
Single 9 ref 0.69 [0.61 , 0.77]
Multiple 6 −0.04 0.65 [0.59 , 0.71]
Population based 1 −0.32 0.37 [0.08 , 0.79]
ESRS
Single 4 ref 0.60 [0.55 , 0.65]
Multiple 7 0.00 0.60 [0.58 , 0.62]
EuroSCORE
Single 12 ref 0.81 [0.78 , 0.83]
Multiple 9 −0.03 0.77 [0.75 , 0.79]
Unclear 1 −0.10 0.70 [0.38 , 0.90]
Framingham
Single 2 ref 0.66 [0.55 , 0.76]
Multiple 6 0.03 0.70 [0.64 , 0.75]
Population based 10 0.01 0.68 [0.63 , 0.72]
Unclear 2 0.02 0.69 [0.57 , 0.78]
FRAX
Single 5 ref 0.68 [0.61 , 0.74]
Multiple 9 −0.05 0.63 [0.59 , 0.67]
Population based 15 −0.01 0.67 [0.64 , 0.70]
Unclear 1 −0.04 0.64 [0.51 , 0.75]
ISS
Single 18 ref 0.88 [0.84 , 0.91]
Multiple 15 −0.05 0.82 [0.77 , 0.87]
Unclear 1 0.00 0.88 [0.66 , 0.96]
MELD
Single 12 ref 0.65 [0.59 , 0.70]
Multiple 2 −0.04 0.61 [0.49 , 0.71]
PSI
Single 14 ref 0.79 [0.75 , 0.83]
Multiple 10 0.01 0.80 [0.75 , 0.84]
RCRI
Single 17 ref 0.70 [0.66 , 0.74]
Multiple 6 −0.05 0.66 [0.59 , 0.72]
SAPS 3
Single 13 ref 0.84 [0.80 , 0.87]
Multiple 14 −0.03 0.81 [0.77 , 0.85]
0.40 0.50 0.60 0.70 0.80 0.90 1.00
C−statistic


Changes made to predictors
N
c diff
c [95%CI]
ABCD2
No 16 ref 0.66 [0.61 , 0.71]
ESRS
No 10 ref 0.60 [0.58 , 0.62]
Unclear 1 0.00 0.60 [0.56 , 0.64]
EuroSCORE
No 13 ref 0.80 [0.77 , 0.82]
Yes 5 −0.02 0.78 [0.74 , 0.81]
Unclear 4 −0.02 0.78 [0.70 , 0.84]
Framingham
No 20 ref 0.68 [0.65 , 0.71]
FRAX
No 20 ref 0.66 [0.63 , 0.69]
Yes 10 −0.01 0.65 [0.61 , 0.69]
ISS
No 21 ref 0.86 [0.82 , 0.89]
Unclear 13 0.00 0.85 [0.80 , 0.89]
MELD
No 12 ref 0.64 [0.59 , 0.69]
Unclear 2 −0.05 0.59 [0.44 , 0.73]
PSI
No 13 ref 0.80 [0.76 , 0.84]
Yes 5 0.00 0.81 [0.74 , 0.86]
Unclear 6 −0.03 0.77 [0.71 , 0.83]
RCRI
No 17 ref 0.70 [0.66 , 0.74]
Yes 2 0.01 0.71 [0.60 , 0.80]
Unclear 4 −0.06 0.64 [0.54 , 0.72]
SAPS 3
No 26 ref 0.83 [0.80 , 0.85]
Unclear 1 −0.13 0.70 [0.48 , 0.85]
0.40 0.50 0.60 0.70 0.80 0.90 1.00
C−statistic


Comparability of outcome definition
N
c diff
c [95%CI]
ABCD2
Yes 8 ref 0.69 [0.63 , 0.74]
No 3 −0.10 0.59 [0.49 , 0.68]
Unclear 5 −0.02 0.67 [0.57 , 0.75]
ESRS
Yes 3 ref 0.60 [0.57 , 0.63]
No 8 0.00 0.60 [0.57 , 0.63]
EuroSCORE
Yes 9 ref 0.80 [0.77 , 0.83]
No 11 −0.03 0.78 [0.76 , 0.80]
Unclear 2 0.03 0.83 [0.74 , 0.90]
Framingham
Yes 11 ref 0.69 [0.65 , 0.73]
No 4 −0.01 0.68 [0.61 , 0.74]
Unclear 5 −0.04 0.66 [0.59 , 0.72]
FRAX
Yes 16 ref 0.66 [0.63 , 0.69]
No 13 −0.01 0.65 [0.62 , 0.68]
Unclear 1 0.07 0.73 [0.58 , 0.84]
ISS
Yes 7 ref 0.90 [0.85 , 0.93]
No 19 −0.08 0.82 [0.78 , 0.86]
Unclear 8 −0.02 0.88 [0.83 , 0.92]
MELD
Yes 14 ref 0.64 [0.59 , 0.68]
PSI
Yes 5 ref 0.77 [0.70 , 0.83]
No 18 0.03 0.80 [0.77 , 0.83]
Unclear 1 0.00 0.77 [0.59 , 0.89]
RCRI
Yes 5 ref 0.76 [0.70 , 0.81]
No 18 −0.09 0.67 [0.63 , 0.70]
SAPS 3
Yes 24 ref 0.82 [0.79 , 0.85]
No 3 0.03 0.86 [0.77 , 0.91]
0.40 0.50 0.60 0.70 0.80 0.90 1.00
C−statistic

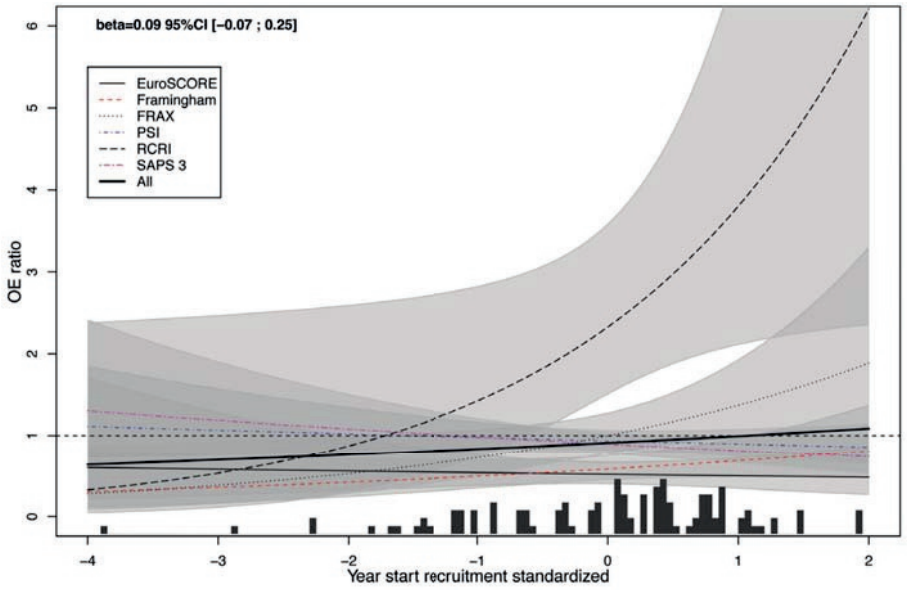

Outcome measurement method
N
c diff
c [95%CI]
ABCD2
Self−reported 3 0.11 0.77 [0.63 , 0.86]
Clinician 6 0.01 0.66 [0.58 , 0.73]
Registry 3 ref 0.65 [0.56 , 0.73]
Unclear 4 −0.04 0.62 [0.49 , 0.73]
ESRS
Self−reported 8 −0.01 0.60 [0.59 , 0.61]
Clinician 1 −0.10 0.51 [0.42 , 0.60]
Registry 2 ref 0.61 [0.56 , 0.65]
EuroSCORE
Self−reported 2 0.02 0.81 [0.74 , 0.86]
Clinician 2 −0.01 0.78 [0.71 , 0.83]
Registry 5 ref 0.79 [0.75 , 0.82]
Unclear 13 0.00 0.79 [0.76 , 0.82]
Framingham
Self−reported 12 −0.07 0.67 [0.64 , 0.70]
Clinician 2 −0.12 0.62 [0.52 , 0.71]
Registry 4 ref 0.74 [0.69 , 0.78]
Unclear 2 −0.07 0.67 [0.56 , 0.76]
FRAX
Self−reported 18 −0.02 0.65 [0.62 , 0.68]
Clinician 1 −0.06 0.62 [0.48 , 0.73]
Registry 8 ref 0.67 [0.63 , 0.72]
Unclear 3 −0.01 0.66 [0.58 , 0.73]
ISS
Clinician 4 0.01 0.85 [0.75 , 0.92]
Registry 13 ref 0.85 [0.79 , 0.89]
Unclear 17 0.02 0.86 [0.82 , 0.90]
MELD
Self−reported 1 0.28 0.95 [0.01 , 1.00]
Registry 3 ref 0.67 [0.56 , 0.76]
Unclear 10 −0.04 0.63 [0.57 , 0.68]
PSI
Self−reported 6 0.01 0.81 [0.75 , 0.85]
Clinician 2 0.05 0.85 [0.76 , 0.91]
Registry 9 ref 0.79 [0.75 , 0.83]
Unclear 7 −0.03 0.77 [0.71 , 0.82]
RCRI
Self−reported 1 −0.09 0.62 [0.43 , 0.78]
Clinician 10 −0.01 0.70 [0.64 , 0.75]
Registry 6 ref 0.71 [0.65 , 0.77]
Unclear 6 −0.07 0.64 [0.55 , 0.73]
SAPS 3
Self−reported 2 0.02 0.83 [0.72 , 0.91]
Clinician 6 0.03 0.85 [0.79 , 0.89]
Registry 9 ref 0.81 [0.76 , 0.86]
Unclear 10 0.00 0.82 [0.77 , 0.86]
0.40 0.50 0.60 0.70 0.80 0.90 1.00
C−statistic

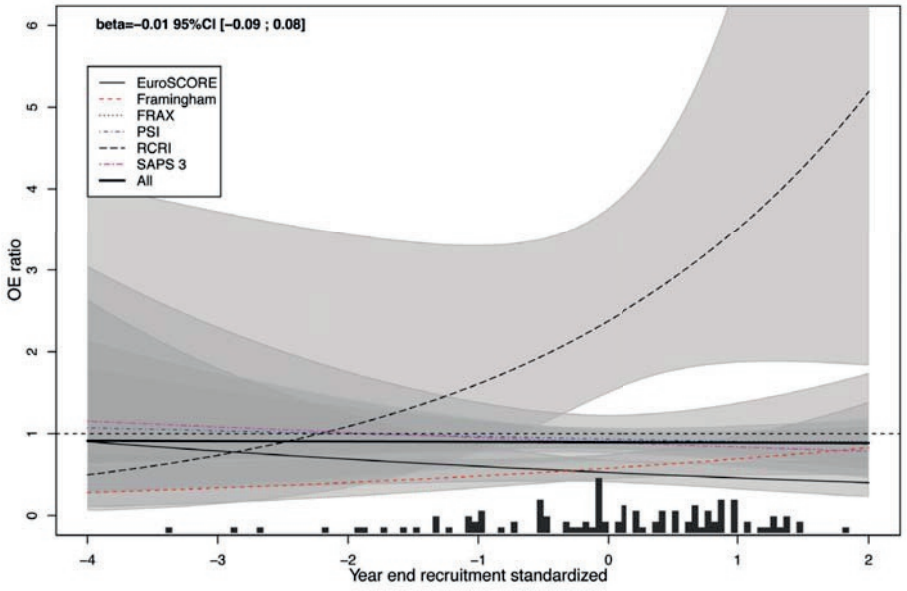


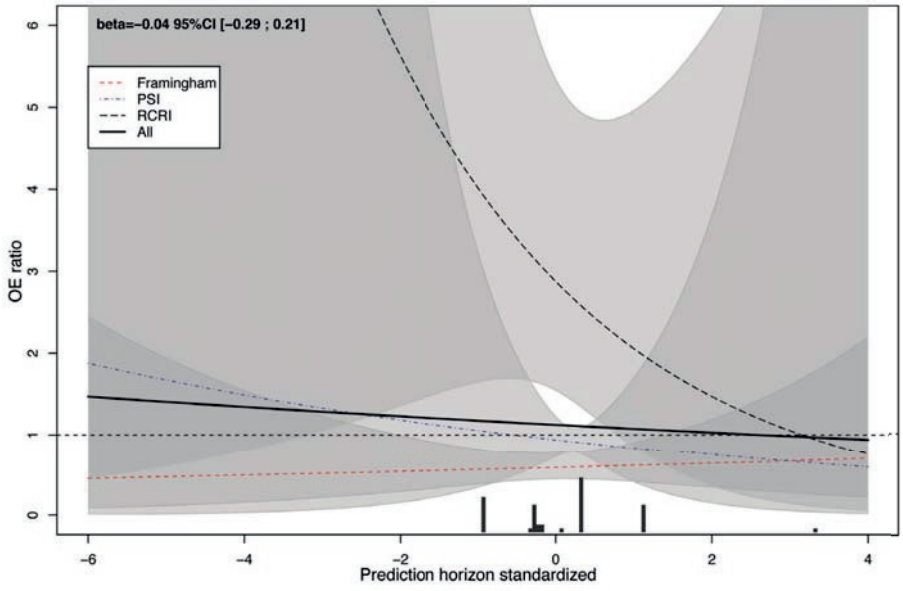


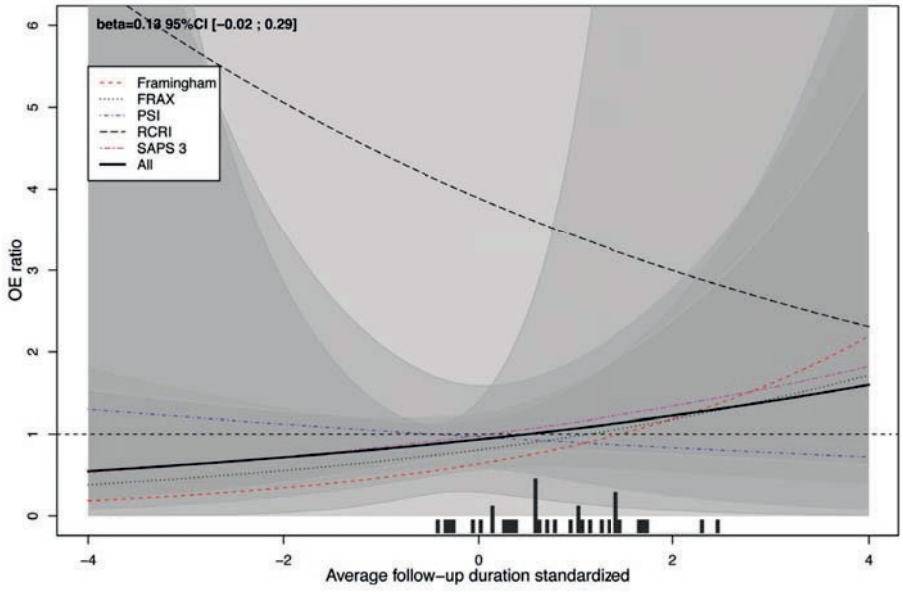
C-statistic for categories of study characteristics, pooled using univariable meta-regression analyses per systematic review. N represents the number of external validation studies in a specific category. C diff represents the difference in c-statistic with regard to a reference category (indicated with 'ref'). Dev: development, val: validation, incr: incremental value, pts: patients. *Models contain age as predictor

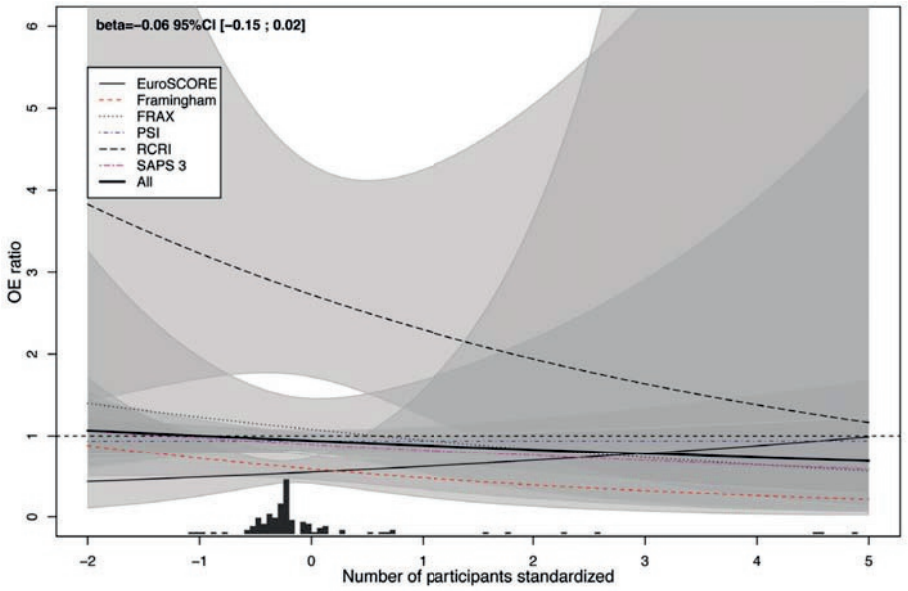
Figure S3 Associations between continuous variables and OE ratio

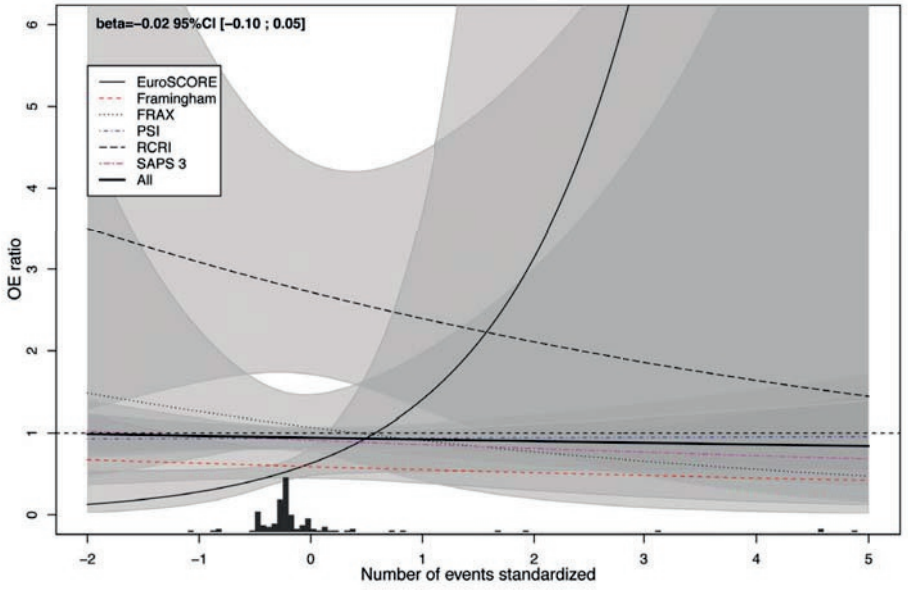


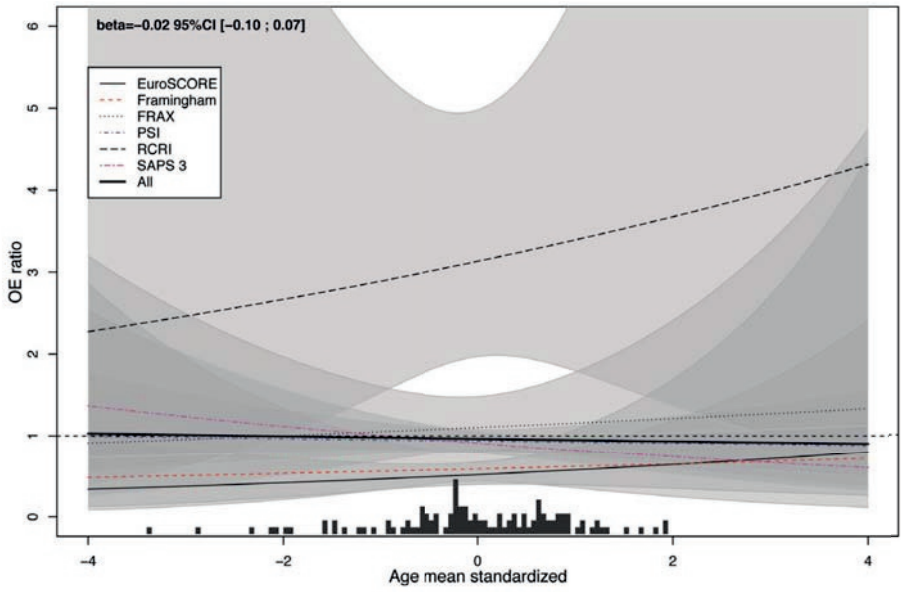


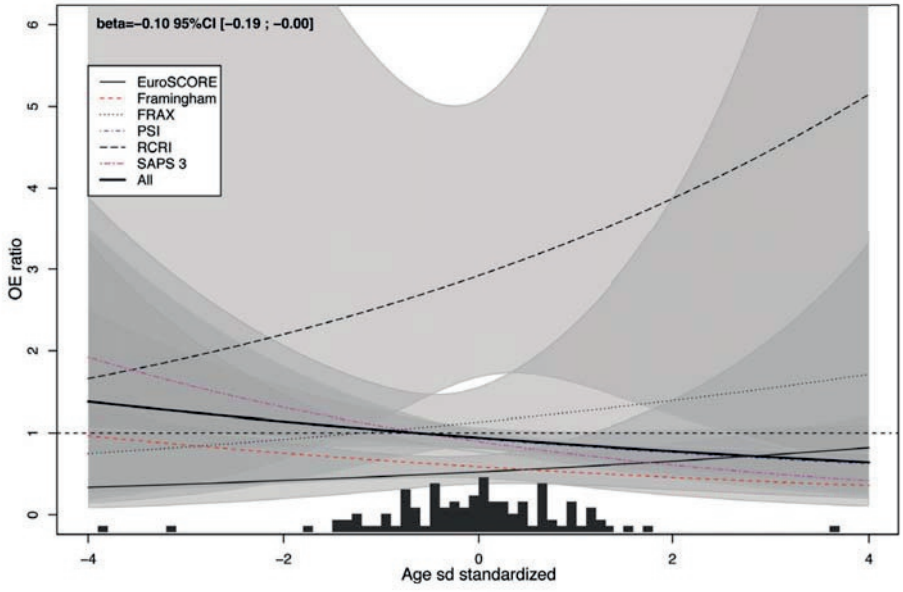

beta=−0.04 95%CI [−0.29 ; 0.21]
Framingham
PSI
RCRI
All
OE ratio
Prediction horizon standardized


beta=0.13 95%CI [−0.02 ; 0.29]
Framingham
FRAX
PSI
RCRI
SAPS 3
All
OE ratio
Average follow−up duration standardized


beta=−0.06 95%CI [−0.15 ; 0.02]
EuroSCORE
Framingham
FRAX
PSI
RCRI
SAPS 3
All
OE ratio
Number of participants standardized


beta=−0.02 95%CI [−0.10 ; 0.05]
EuroSCORE
Framingham
FRAX
PSI
RCRI
SAPS 3
All
OE ratio
Number of events standardized


beta=−0.02 95%CI [−0.10 ; 0.07]
EuroSCORE
Framingham
FRAX
PSI
RCRI
SAPS 3
All
OE ratio
0
1
2
3
4
5
6
−4
−2
0
2
4
Age mean standardized


beta=−0.10 95%CI [−0.19 ; −0.00]
EuroSCORE
Framingham
FRAX
PSI
RCRI
SAPS 3
All
OE ratio
0
1
2
3
4
5
6
−4
−2
0
2
4
Age sd standardized

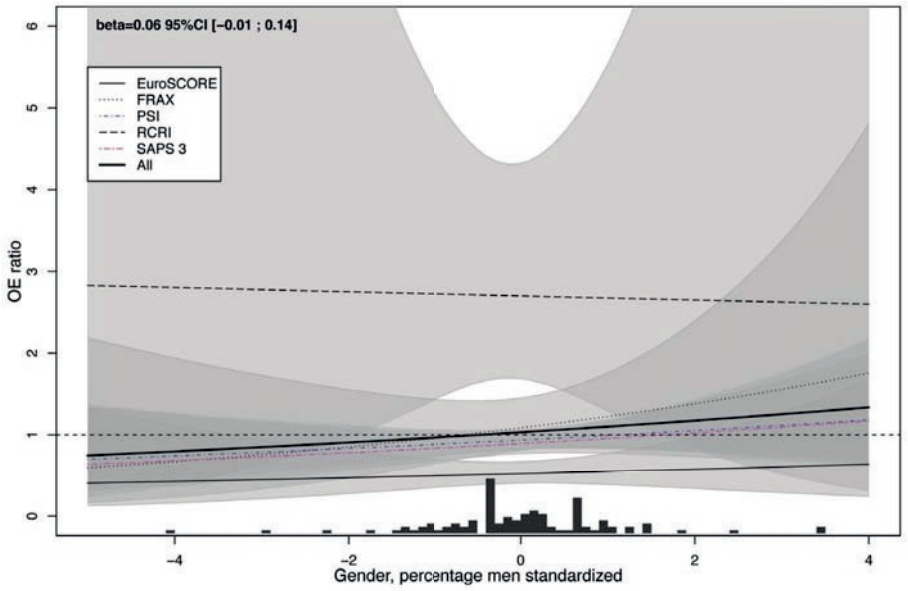
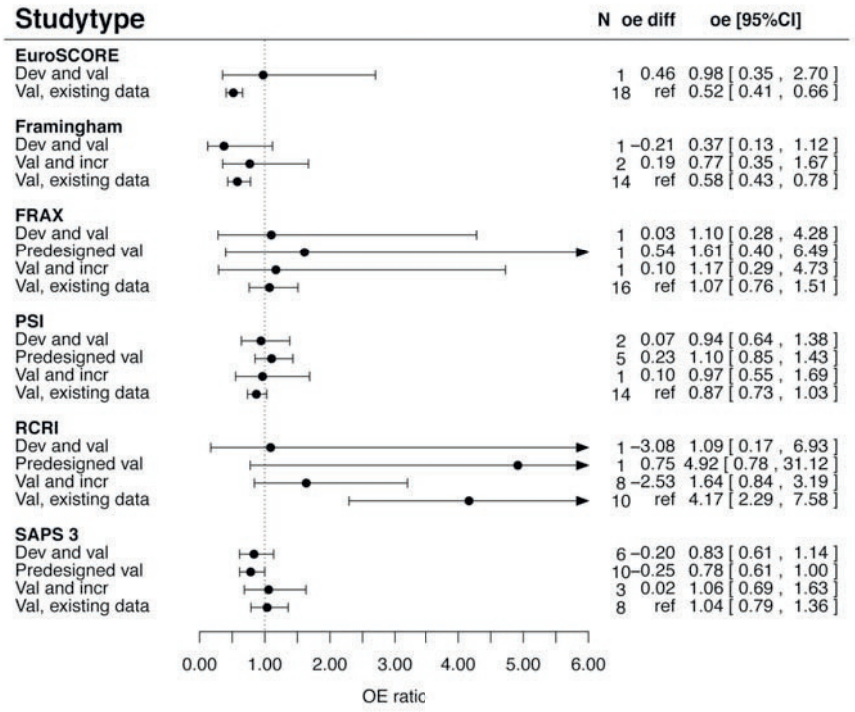
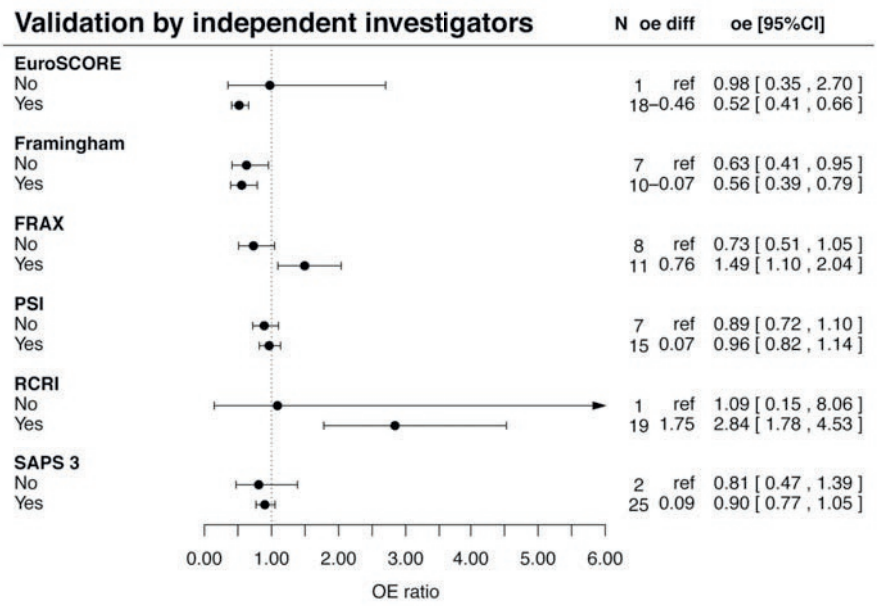
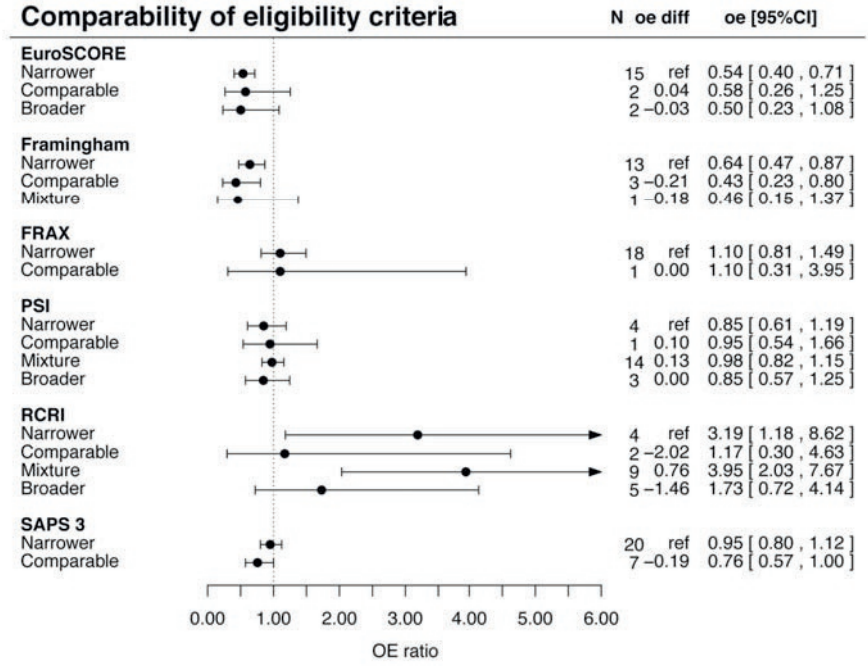
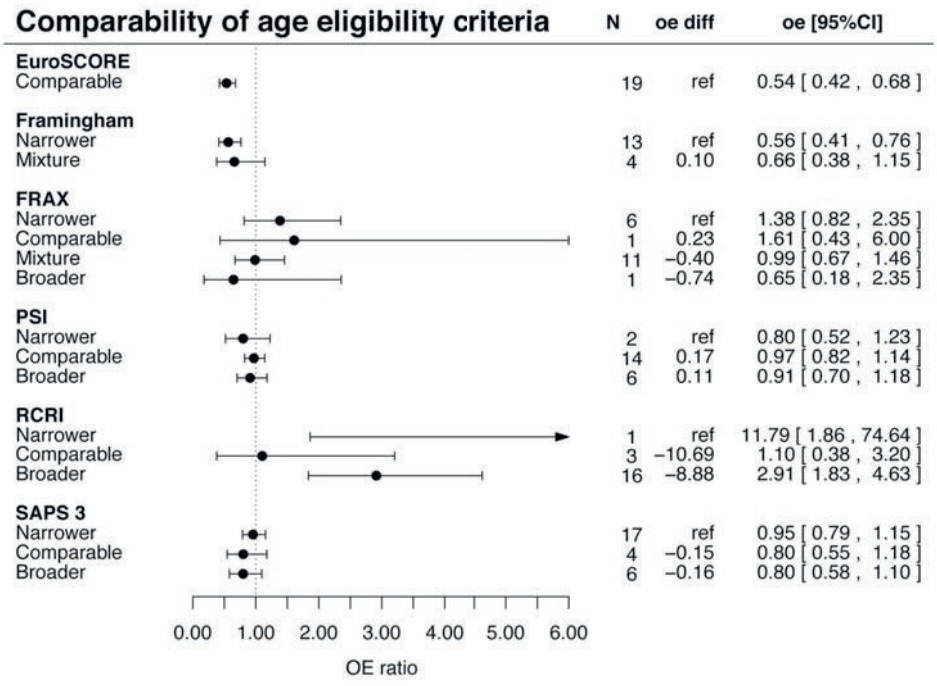


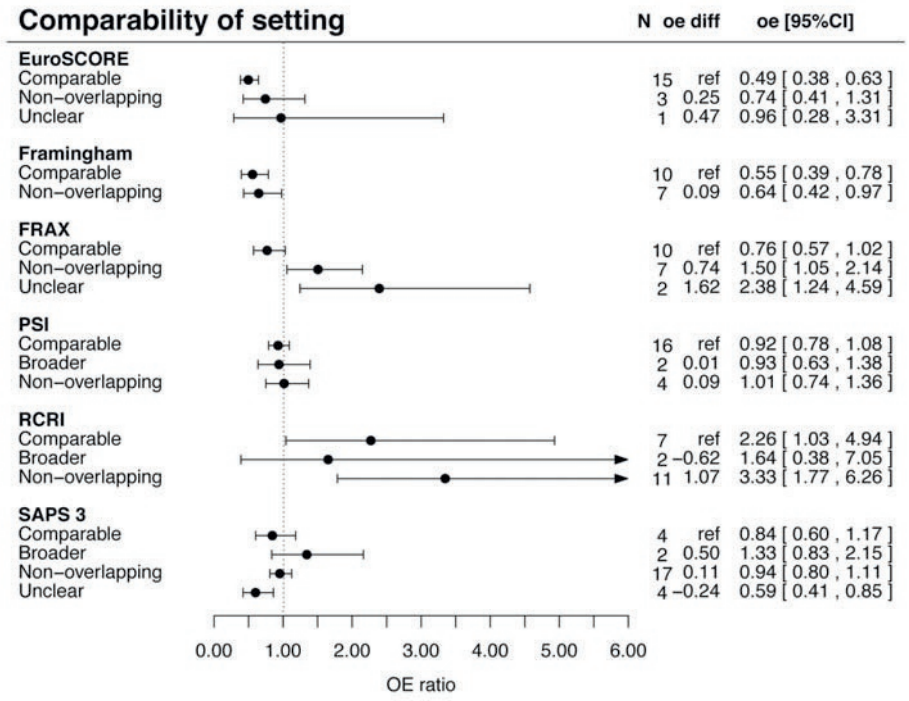
Figure S4: OE ratio in categories of study characteristics within each systematic review

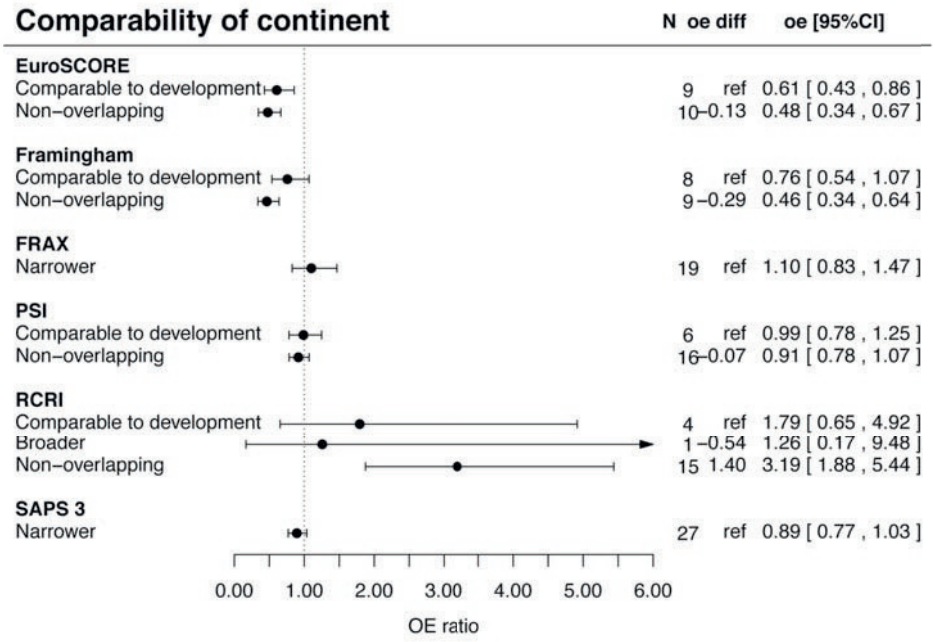


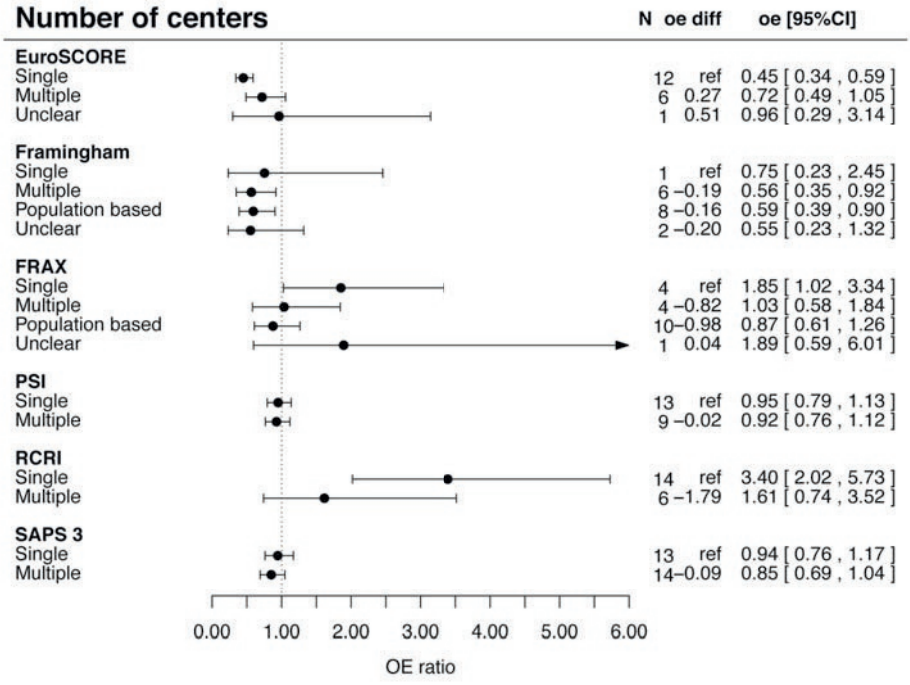

Validation by independent investigators
N oe diff oe [95%CI]
EuroSCORE
No 1 ref 0.98 [0.35 , 2.70]
Yes 18 −0.46 0.52 [0.41 , 0.66]
Framingham
No 7 ref 0.63 [0.41 , 0.95]
Yes 10 −0.07 0.56 [0.39 , 0.79]
FRAX
No 8 ref 0.73 [0.51 , 1.05]
Yes 11 0.76 1.49 [1.10 , 2.04]
PSI
No 7 ref 0.89 [0.72 , 1.10]
Yes 15 0.07 0.96 [0.82 , 1.14]
RCRI
No 1 ref 1.09 [0.15 , 8.06]
Yes 19 1.75 2.84 [1.78 , 4.53]
SAPS 3
No 2 ref 0.81 [0.47 , 1.39]
Yes 25 0.09 0.90 [0.77 , 1.05]
0.00 1.00 2.00 3.00 4.00 5.00 6.00
OE ratio

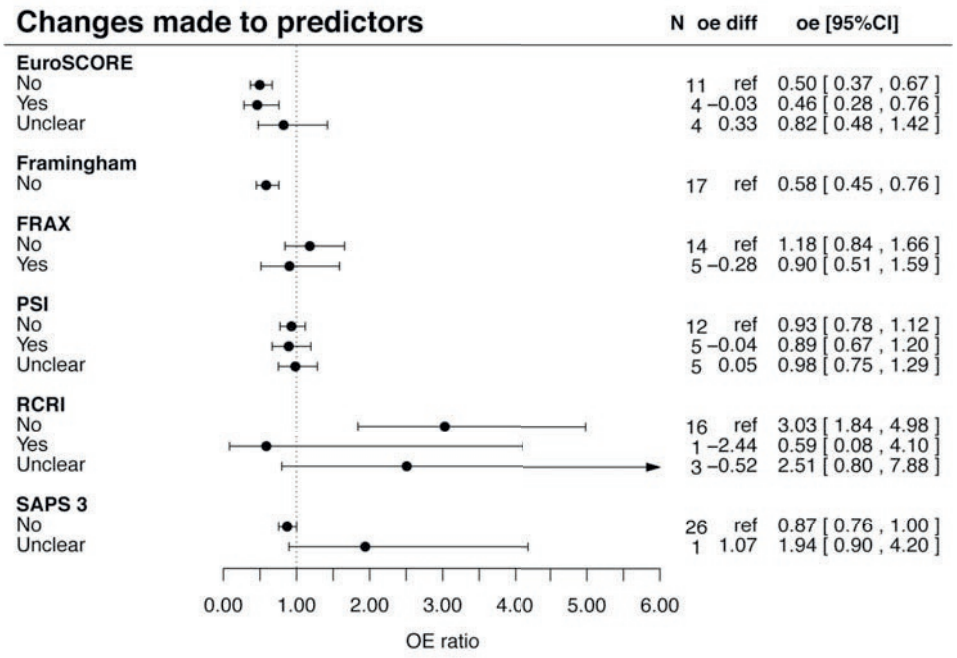

Comparability of eligibility criteria
N oe diff oe [95%CI]
EuroSCORE
Narrower 15 ref 0.54 [0.40 , 0.71]
Comparable 2 0.04 0.58 [0.26 , 1.25]
Broader 2 −0.03 0.50 [0.23 , 1.08]
Framingham
Narrower 13 ref 0.64 [0.47 , 0.87]
Comparable 3 −0.21 0.43 [0.23 , 0.80]
Mixture 1 −0.18 0.46 [0.15 , 1.37]
FRAX
Narrower 18 ref 1.10 [0.81 , 1.49]
Comparable 1 0.00 1.10 [0.31 , 3.95]
PSI
Narrower 4 ref 0.85 [0.61 , 1.19]
Comparable 1 0.10 0.95 [0.54 , 1.66]
Mixture 14 0.13 0.98 [0.82 , 1.15]
Broader 3 0.00 0.85 [0.57 , 1.25]
RCRI
Narrower 4 ref 3.19 [1.18 , 8.62]
Comparable 2 −2.02 1.17 [0.30 , 4.63]
Mixture 9 0.76 3.95 [2.03 , 7.67]
Broader 5 −1.46 1.73 [0.72 , 4.14]
SAPS 3
Narrower 20 ref 0.95 [0.80 , 1.12]
Comparable 7 −0.19 0.76 [0.57 , 1.00]
0.00 1.00 2.00 3.00 4.00 5.00 6.00
OE ratio

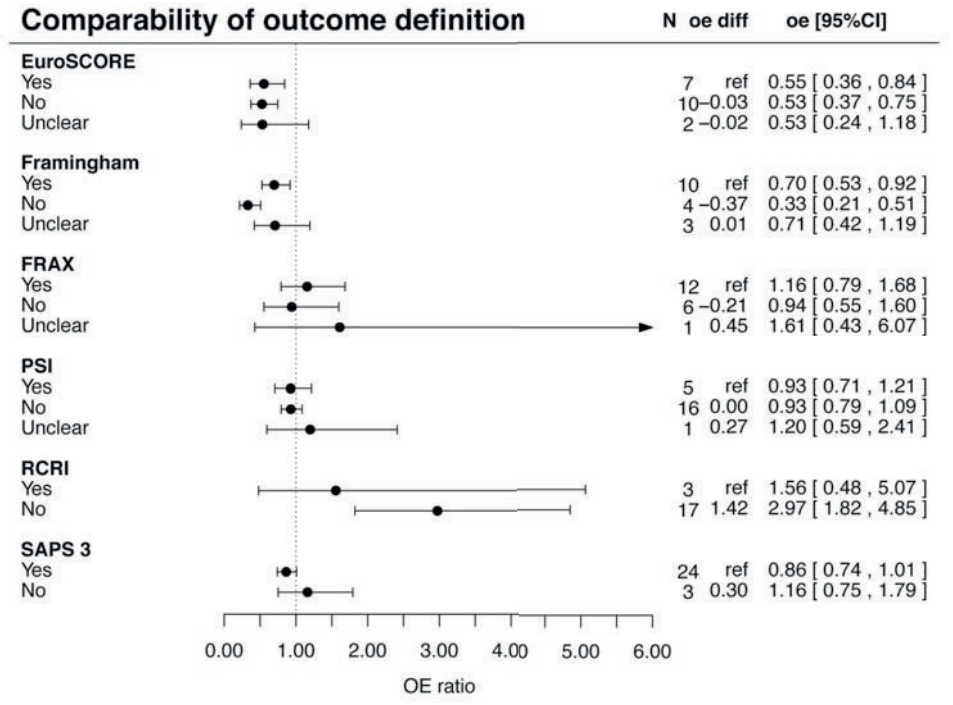

Comparability of age eligibility criteria
N
oe diff
oe [95%CI]
EuroSCORE
Comparable 19 ref 0.54 [0.42 , 0.68]
Framingham
Narrower 13 ref 0.56 [0.41 , 0.76]
Mixture 4 0.10 0.66 [0.38 , 1.15]
FRAX
Narrower 6 ref 1.38 [0.82 , 2.35]
Comparable 1 0.23 1.61 [0.43 , 6.00]
Mixture 11 −0.40 0.99 [0.67 , 1.46]
Broader 1 −0.74 0.65 [0.18 , 2.35]
PSI
Narrower 2 ref 0.80 [0.52 , 1.23]
Comparable 14 0.17 0.97 [0.82 , 1.14]
Broader 6 0.11 0.91 [0.70 , 1.18]
RCRI
Narrower 1 ref 11.79 [1.86 , 74.64]
Comparable 3 −10.69 1.10 [0.38 , 3.20]
Broader 16 −8.88 2.91 [1.83 , 4.63]
SAPS 3
Narrower 17 ref 0.95 [0.79 , 1.15]
Comparable 4 −0.15 0.80 [0.55 , 1.18]
Broader 6 −0.16 0.80 [0.58 , 1.10]
0.00 1.00 2.00 3.00 4.00 5.00 6.00
OE ratio

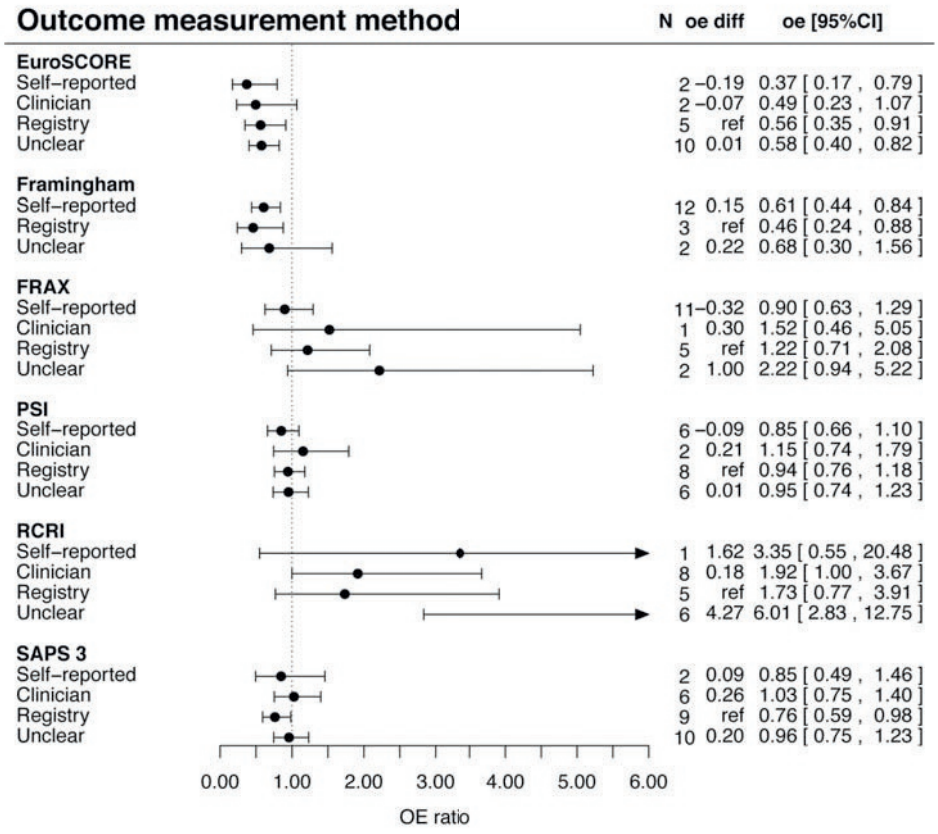

Comparability of setting
N
oe diff
oe [95%CI]
EuroSCORE
Comparable 15 ref 0.49 [0.38 , 0.63]
Non−overlapping 3 0.25 0.74 [0.41 , 1.31]
Unclear 1 0.47 0.96 [0.28 , 3.31]
Framingham
Comparable 10 ref 0.55 [0.39 , 0.78]
Non−overlapping 7 0.09 0.64 [0.42 , 0.97]
FRAX
Comparable 10 ref 0.76 [0.57 , 1.02]
Non−overlapping 7 0.74 1.50 [1.05 , 2.14]
Unclear 2 1.62 2.38 [1.24 , 4.59]
PSI
Comparable 16 ref 0.92 [0.78 , 1.08]
Broader 2 0.01 0.93 [0.63 , 1.38]
Non−overlapping 4 0.09 1.01 [0.74 , 1.36]
RCRI
Comparable 7 ref 2.26 [1.03 , 4.94]
Broader 2 −0.62 1.64 [0.38 , 7.05]
Non−overlapping 11 1.07 3.33 [1.77 , 6.26]
SAPS 3
Comparable 4 ref 0.84 [0.60 , 1.17]
Broader 2 0.50 1.33 [0.83 , 2.15]
Non−overlapping 17 0.11 0.94 [0.80 , 1.11]
Unclear 4 −0.24 0.59 [0.41 , 0.85]
0.00 1.00 2.00 3.00 4.00 5.00 6.00
OE ratio

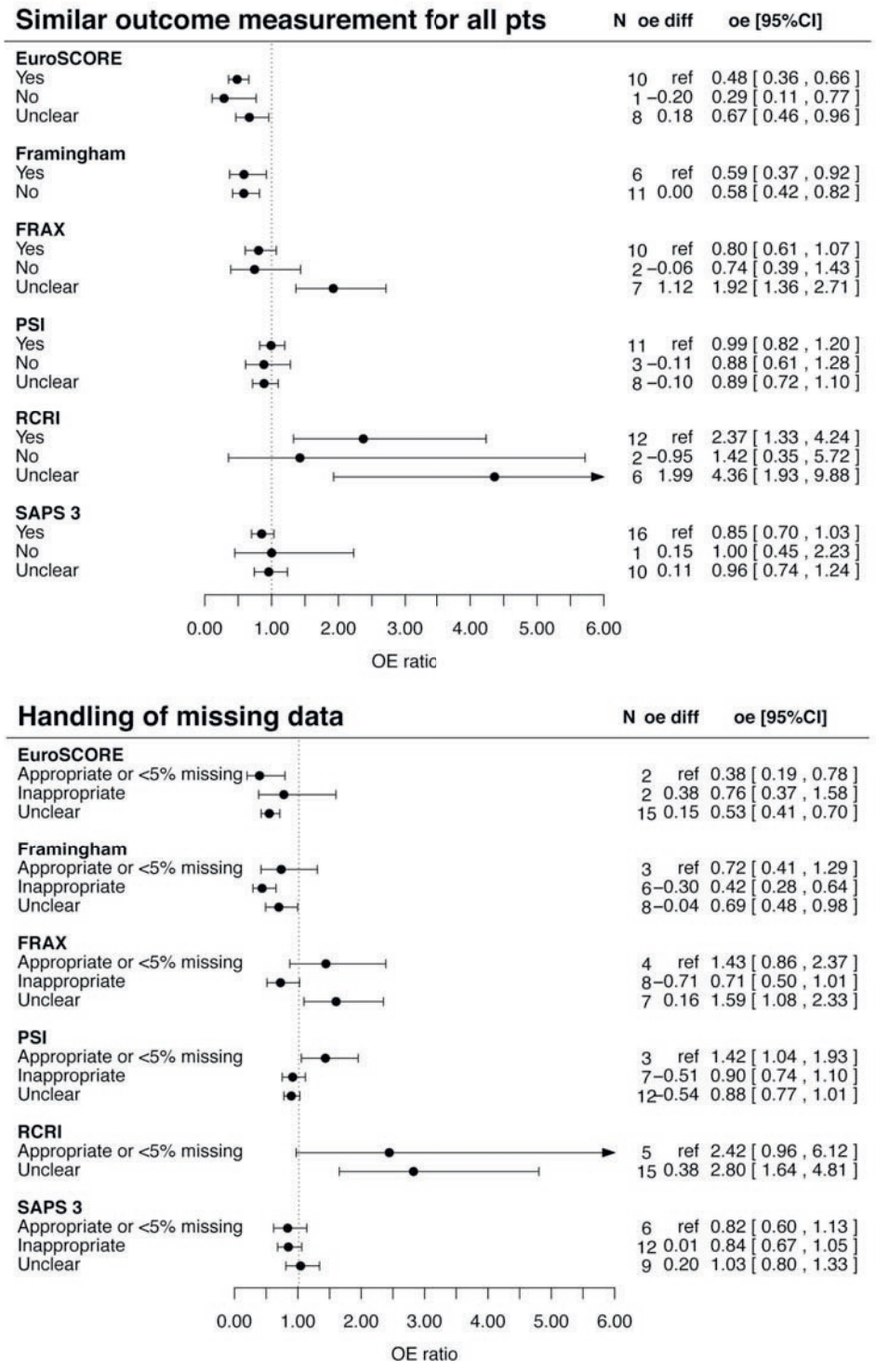

Comparability of continent
N oe diff oe [95%CI]
EuroSCORE
Comparable to development 9 ref 0.61 [0.43 , 0.86]
Non−overlapping 10 −0.13 0.48 [0.34 , 0.67]
Framingham
Comparable to development 8 ref 0.76 [0.54 , 1.07]
Non−overlapping 9 −0.29 0.46 [0.34 , 0.64]
FRAX
Narrower 19 ref 1.10 [0.83 , 1.47]
PSI
Comparable to development 6 ref 0.99 [0.78 , 1.25]
Non−overlapping 16 −0.07 0.91 [0.78 , 1.07]
RCRI
Comparable to development 4 ref 1.79 [0.65 , 4.92]
Broader 1 −0.54 1.26 [0.17 , 9.48]
Non−overlapping 15 1.40 3.19 [1.88 , 5.44]
SAPS 3
Narrower 27 ref 0.89 [0.77 , 1.03]
0.00 1.00 2.00 3.00 4.00 5.00 6.00
OE ratio


Number of centers
N oe diff oe [95%CI]
EuroSCORE
Single 12 ref 0.45 [0.34 , 0.59]
Multiple 6 0.27 0.72 [0.49 , 1.05]
Unclear 1 0.51 0.96 [0.29 , 3.14]
Framingham
Single 1 ref 0.75 [0.23 , 2.45]
Multiple 6 −0.19 0.56 [0.35 , 0.92]
Population based 8 −0.16 0.59 [0.39 , 0.90]
Unclear 2 −0.20 0.55 [0.23 , 1.32]
FRAX
Single 4 ref 1.85 [1.02 , 3.34]
Multiple 4 −0.82 1.03 [0.58 , 1.84]
Population based 10 −0.98 0.87 [0.61 , 1.26]
Unclear 1 0.04 1.89 [0.59 , 6.01]
PSI
Single 13 ref 0.95 [0.79 , 1.13]
Multiple 9 −0.02 0.92 [0.76 , 1.12]
RCRI
Single 14 ref 3.40 [2.02 , 5.73]
Multiple 6 −1.79 1.61 [0.74 , 3.52]
SAPS 3
Single 13 ref 0.94 [0.76 , 1.17]
Multiple 14 −0.09 0.85 [0.69 , 1.04]
0.00 1.00 2.00 3.00 4.00 5.00 6.00
OE ratio


Changes made to predictors
N oe diff oe [95%CI]
EuroSCORE
No 11 ref 0.50 [0.37 , 0.67]
Yes 4 −0.03 0.46 [0.28 , 0.76]
Unclear 4 0.33 0.82 [0.48 , 1.42]
Framingham
No 17 ref 0.58 [0.45 , 0.76]
FRAX
No 14 ref 1.18 [0.84 , 1.66]
Yes 5 −0.28 0.90 [0.51 , 1.59]
PSI
No 12 ref 0.93 [0.78 , 1.12]
Yes 5 −0.04 0.89 [0.67 , 1.20]
Unclear 5 0.05 0.98 [0.75 , 1.29]
RCRI
No 16 ref 3.03 [1.84 , 4.98]
Yes 1 −2.44 0.59 [0.08 , 4.10]
Unclear 3 −0.52 2.51 [0.80 , 7.88]
SAPS 3
No 26 ref 0.87 [0.76 , 1.00]
Unclear 1 1.07 1.94 [0.90 , 4.20]
0.00 1.00 2.00 3.00 4.00 5.00 6.00
OE ratio


Comparability of outcome definition
N oe diff oe [95%CI]
EuroSCORE
Yes 7 ref 0.55 [0.36 , 0.84]
No 10 −0.03 0.53 [0.37 , 0.75]
Unclear 2 −0.02 0.53 [0.24 , 1.18]
Framingham
Yes 10 ref 0.70 [0.53 , 0.92]
No 4 −0.37 0.33 [0.21 , 0.51]
Unclear 3 0.01 0.71 [0.42 , 1.19]
FRAX
Yes 12 ref 1.16 [0.79 , 1.68]
No 6 −0.21 0.94 [0.55 , 1.60]
Unclear 1 0.45 1.61 [0.43 , 6.07]
PSI
Yes 5 ref 0.93 [0.71 , 1.21]
No 16 0.00 0.93 [0.79 , 1.09]
Unclear 1 0.27 1.20 [0.59 , 2.41]
RCRI
Yes 3 ref 1.56 [0.48 , 5.07]
No 17 1.42 2.97 [1.82 , 4.85]
SAPS 3
Yes 24 ref 0.86 [0.74 , 1.01]
No 3 0.30 1.16 [0.75 , 1.79]
0.00 1.00 2.00 3.00 4.00 5.00 6.00
OE ratio


Outcome measurement method
N oe diff oe [95%CI]
EuroSCORE
Self−reported 2 −0.19 0.37 [0.17 , 0.79]
Clinician 2 −0.07 0.49 [0.23 , 1.07]
Registry 5 ref 0.56 [0.35 , 0.91]
Unclear 10 0.01 0.58 [0.40 , 0.82]
Framingham
Self−reported 12 0.15 0.61 [0.44 , 0.84]
Registry 3 ref 0.46 [0.24 , 0.88]
Unclear 2 0.22 0.68 [0.30 , 1.56]
FRAX
Self−reported 11 −0.32 0.90 [0.63 , 1.29]
Clinician 1 0.30 1.52 [0.46 , 5.05]
Registry 5 ref 1.22 [0.71 , 2.08]
Unclear 2 1.00 2.22 [0.94 , 5.22]
PSI
Self−reported 6 −0.09 0.85 [0.66 , 1.10]
Clinician 2 0.21 1.15 [0.74 , 1.79]
Registry 8 ref 0.94 [0.76 , 1.18]
Unclear 6 0.01 0.95 [0.74 , 1.23]
RCRI
Self−reported 1 1.62 3.35 [0.55 , 20.48]
Clinician 8 0.18 1.92 [1.00 , 3.67]
Registry 5 ref 1.73 [0.77 , 3.91]
Unclear 6 4.27 6.01 [2.83 , 12.75]
SAPS 3
Self−reported 2 0.09 0.85 [0.49 , 1.46]
Clinician 6 0.26 1.03 [0.75 , 1.40]
Registry 9 ref 0.76 [0.59 , 0.98]
Unclear 10 0.20 0.96 [0.75 , 1.23]
0.00
1.00
2.00
3.00
4.00
5.00
6.00
OE ratio



OE ratio for categories of study characteristics, pooled using univariable meta-regression analyses per systematic review. N represents the number of external validation studies in a specific category. OE diff represents the difference in OE ratio with regard to a reference category (indicated with 'ref'). Dev: development, val: validation, incr: incremental value, pts: patients.

References

1. Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res* 2017:962280217705678.
2. Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: asymptotic methods and evaluation. *Stat Med* 2006;25(4):559-73.
3. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29-36.
4. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;356:i6460.
5. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* 2014;14:25.
6. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2015.
7. R: A language and environment for statistical computing [program]. Vienna, Austria: R Foundation for Statistical Computing, 2016.
8. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 2010;36(3):1-48.
9. Gasparrini A, Armstrong B, Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Stat Med* 2012;31(29):3821-39.
10. Debray TP. Metamisc: Diagnostic and Prognostic Meta-Analysis. 2017.
11. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw* 2015;67(1).

Chapter 9

General discussion

This thesis aimed to provide guidance on how to perform systematic reviews and meta-analyses of prediction model studies. We applied the developed guidance on studies in the field of cardiovascular disease (CVD), thereby identifying generic issues that require further attention in future methodological and empirical studies.

Lessons learnt

- We presented guidance for systematically reviewing and meta-analysing existing evidence regarding diagnostic and prognostic models (Chapter 2).
- We identified more than 300 prognostic models for the prediction of CVD in the general population. Due to methodological shortcomings, incomplete presentation, lack of external validation studies and the absence of model impact studies, the usefulness in clinical practice of most models is unclear (Chapter 3).
- In a meta-analysis of three frequently advocated models for the prediction of coronary heart disease (CHD) or CVD (Framingham Wilson, Framingham ATP III and Pooled Cohort Equations (PCE)) the discriminative performance of these models was comparable. All three models overestimated the risk of developing CHD or CVD, especially in higher risk populations (Chapter 4).
- In a large multicentre European cohort, we found limited incremental value of biomarkers over traditional predictors for the prediction of 10 year risk of CHD (Chapter 5).
- The majority of studies in which prognostic models for CVD were developed or validated did not take into account the use of treatment that may lower CVD risks during follow-up. In addition, information about treatment use was infrequently reported (Chapter 6).
- In a systematic review, we showed that more than half of the items considered essential for reporting according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement were not fully addressed. Essential information for using a model for individual risk prediction was frequently incomplete (Chapter 7).
- Using a meta-epidemiological approach, we showed that variation in the predictive performance of prognostic models is particularly related to variation in study population (i.e. case-mix), and predictor and outcome measurement (Chapter 8).

Future perspectives for cardiovascular risk prediction research

Surprisingly, the majority of the models that we identified for the estimation of future risk of CVD in the general population, were very similar. The aim of such models is to better target prevention strategies to decrease the number of CVD events or to delay the time to a CVD event. We found that most models consisted of the same set of core predictors and predicted similar outcomes. We, therefore, believe that future cardiovascular risk prediction research aimed to target prevention strategies, should not focus on developing new prognostic models, but on the external validation of the available models, especially of models that consist of predictors that can easily be

measured. Such prognostic models are more likely to be used than models with more difficult to measure predictors. Also, we studied the added value of novel biomarkers, which turned out to be limited (Chapter 5).[1,2] This confirms the usefulness of the easy to measure predictors. Preferably, future external validation studies of the more simple CVD risk prediction models should be combined with updating (i.e. adjusting the parameters of a prediction model) to improve their predictive performance,[3-5] since most existing models are overestimating the actual number of CVD events when applied in new settings. It has been argued, that this overestimation could be the result of flaws in the study design or the use of treatment (e.g. lipid lowering and anti-hypertensive drugs) that has changed over time since the development of the models. Such increased treatment use lowers the observed risks of the outcomes, leading to overestimated predicted risks by the previously developed models.[6-8] However, it has also been shown that the use of treatment does not fully explain this overestimation.[9] On an aggregated level in our meta-analytical study (Chapter 3) we were also not able to gain insight into the impact of treatment use on overprediction due to poor reporting thereof in the primary studies. A better and far more efficient way to investigate the role of treatment use on the overestimation in prognostic models and to explore other possible explanations for the observed overestimation, is offered by meta-analyses based on individual participant data (IPD) of published studies. To improve the existing prognostic models, routinely collected data such as electronic health care records, can be used to tailor the models to specific settings and to continuously update them with new data.[10] Methods that take into account treatment-covariate interactions,[11,12] or dynamic prediction modelling methods[13-15] may also improve the performance of models in the field of CVD risk prediction, to solve the challenge of overestimation of risks.

Currently, the prediction horizon of the various models is 5 or 10 years. In most models, age is a major driver of prediction. However, in young people with multiple CVD risk factors the estimated risk of getting an CVD event within 10-years is often low, although they might be at high long-term risk.[16,17] These people could have benefits of preventative strategies, but based on the model predictions they are currently not considered eligible for such treatments.[18,19] A new type of modelling in which lifetime risk is estimated instead of 5- or 10-year risks, combined with information on treatments and their efficacy, may offer a solution to decide who might benefit most from preventative treatment.[18]

The aim, as said, of using CVD risk prediction models is to decrease the number of CVD events or to delay the time to the (next) CVD event. So ultimately, one would like to know the impact of using such prognostic models on physicians' and individuals' behaviour (e.g. treatment prescription and lifestyle changes), and more importantly, the impact on subsequent health outcomes. However, such impact of these CVD risk prediction models has rarely been evaluated. The available studies have shown that using prognostic models increases the prescription of lipid lowering and antihypertensive

medication. The effect of these models on downstream changes in the risk factors for CVD is limited, and there is no evidence on changes in the occurrence of CVD events.[20,21] Once prognostic models with good predictive performance are available, studies with a comparative design should ideally focus on evaluating the effect of using these models (as compared to absence of their use) in practice on the changes in both CVD risk factors and the actual occurrence of CVD events.

Future perspectives for prediction model research in general

If for a certain medical condition multiple prediction models are available, systematic reviews are useful to identify which model is best in which situation (e.g. for which subgroup of patients, for which setting or for which country). Unawareness of existing prediction models and non-reporting of items essential for applying the model in practice, are still major barriers for researchers and users of prediction models. We have developed methods to guide authors of systematic reviews, which may also contribute to the quality of reporting and conduct of primary studies on prediction models. Recently, guidance has become available not only regarding the preferred reporting of prediction model studies (TRIPOD[22,23]), but also on data extraction and critical appraisal of prediction model studies (CHARMS[24]). In addition, a formal tool to assess risk of bias in prediction model studies is about to be published (PROBAST[25]). It is important that (future) researchers are aware of these methods and are going to use them. We aim to improve awareness and implementation of these tools. It has been shown that active multicomponent strategies, that address various target groups, are more effective than passive and single component strategies.[26] Therefore, we plan to provide better education to our (bio)medical students, develop freely available online education materials for researchers, journal editors, peer reviewers, healthcare providers, and students, raise awareness of these tools and methods at scientific conferences and publish further guidance and training.

Poor reporting is an important source of research waste.[27-29] In addition, complete reporting in primary studies is a requirement for informative systematic reviews, notably in the field of prediction models. Unfortunately, crucial information regarding the methods or results is often not reported in prediction model studies. Therefore, many studies cannot be included in meta-analyses, which could lead to bias in pooled effect estimates. In 2015, the TRIPOD statement for the reporting of primary prediction studies[22,23] was published. To facilitate the uptake of the TRIPOD statement, to ensure that all future published studies can be critically appraised and correctly interpreted by other researchers, and to make people aware of the importance of transparent and complete reporting of prediction research, we have started to work with the EQUATOR network (https://www.equator-network.org/). In addition, we will implement our methods within Cochrane, an internationally leading organization for systematic reviews. Because methods advocated by Cochrane are often considered to be the 'gold standard' we are

confident that many other medical journals, peer reviewers and researchers will follow our recommendations.

Meta-epidemiological research of prognostic model studies turned out to be challenging due to heterogeneity between primary studies. For example, we found much variation in the way predictors were measured, outcomes were defined and participants were selected for inclusion in the various studies. Researchers of primary studies should pay more attention to these factors in order to better harmonize studies, which will enable more sound meta-epidemiological studies that can better focus on true characteristics of the design and analyses, such as sample size and handling of missing data. To date, insight into sources of heterogeneity of prediction model performance appears to come mainly from simulation studies. However, this insight should be confirmed in meta-epidemiological studies. IPD meta-analyses can be used to better investigate design related biases as this allows the calculation of various performance measures that are less sensitive to variations in study population, such as the case-mix corrected c-statistic and calibration slope.[30-32]

Concluding remarks

We have developed methods for systematic reviews and meta-analyses of prediction models, thereby encountering several challenges in the design, methodological conduct, reporting, and research focus of primary prediction model studies. We believe it is time for a change and provided several solutions to overcome main barriers for implementing our recommendations and methods. Developing education material for several stakeholders will be one of the cornerstones, implemented by international organizations like the EQUATOR-network and Cochrane.

References

1. Ferket BS, van Kempen BJ, Hunink MG, Agarwal I, Kavousi M, Franco OH, et al. Predictive value of updating Framingham risk scores with novel risk markers in the U.S. general population. *PLoS One* 2014;9(2):e88312.
2. van der Meer MG, van der Graaf Y, Schuit E, Peelen LM, Verschuren WM, Boer JM, et al. Added Value of Female-Specific Factors Beyond Traditional Predictors for Future Cardiovascular Disease. *J Am Coll Cardiol* 2016;67(17):2084-6.
3. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;61(1):76-86.
4. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol* 2008;61(11):1085-94.
5. Nieboer D, Vergouwe Y, Ankerst DP, Roobol MJ, Steyerberg EW. Improving prediction models with new markers: a comparison of updating strategies. *BMC Med Res Methodol* 2016;16(1):128.
6. Goff DC, Jr., D'Agostino RB, Sr., Pencina M, Lloyd-Jones DM. Calibration and Discrimination Among Multiple Cardiovascular Risk Scores in a Modern Multiethnic Cohort. *Ann Intern Med* 2015;163(1):68.
7. Muntner P, Safford MM, Cushman M, Howard G. Comment on the reports of over-estimation of ASCVD risk using the 2013 AHA/ACC risk equation. *Circulation* 2014;129(2):266-7.
8. Spence JD. Statins and ischemic stroke. *JAMA* 2014;312(7):749-50.
9. Cook NR, Ridker PM. Further insight into the cardiovascular risk calculator: the roles of statins, revascularizations, and underascertainment in the Women's Health Study. *JAMA* 2014;174(12):1964-71.
10. Vergouwe Y, Nieboer D, Oostenbrink R, Debray TP, Murray GD, Kattan MW, et al. A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med* 2016.
11. van Klaveren D, Vergouwe Y, Farooq V, Serruys PW, Steyerberg EW. Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. *J Clin Epidemiol* 2015;68(11):1366-74.
12. Pajouheshnia R, Peelen LM, Moons KGM, Reitsma JB, Groenwold RHH. Accounting for treatment use when validating a prognostic model: a simulation study. *BMC Med Res Methodol* 2017;17(1):103.
13. Nicolaie MA, van Houwelingen JC, de Witte TM, Putter H. Dynamic prediction by landmarking in competing risks. *Stat Med* 2013;32(12):2031-47.
14. Teramukai S, Okuda Y, Miyazaki S, Kawamori R, Shirayama M, Teramoto T. Dynamic prediction model and risk assessment chart for cardiovascular disease based on on-treatment blood pressure and baseline risk factors. *Hypertens Res* 2016;39(2):113-8.

15. Akbarov A, Williams R, Brown B, Mamas M, Peek N, Buchan I, et al. A Two-stage Dynamic Model to Enable Updating of Clinical Risk Prediction from Longitudinal Health Record Data: Illustrated with Kidney Function. *Stud Health Technol Inform* 2015;216:696-700.
16. Berry JD, Dyer A, Cai X, Garside DB, Ning H, Thomas A, et al. Lifetime risks of cardiovascular disease. *N Engl J Med* 2012;366(4):321-9.
17. Leening MJG, Cook NR, Ridker PM. Should we reconsider the role of age in treatment allocation for primary prevention of cardiovascular disease? *Eur Heart J* 2017;38(20):1542-47.
18. Dorresteijn JA, Kaasenbrood L, Cook NR, van Kruijsdijk RC, van der Graaf Y, Visseren FL, et al. How to translate clinical trial results into gain in healthy life expectancy for individual patients. *BMJ* 2016;352:i1548.
19. Ferket BS, van Kempen BJ, Heeringa J, Spronk S, Fleischmann KE, Nijhuis RL, et al. Personalized prediction of lifetime benefits with statin therapy for asymptomatic individuals: a modeling study. *PLoS Med* 2012;9(12):e1001361.
20. Usher-Smith JA, Silarova B, Schuit E, Gm Moons K, Griffin SJ. Impact of provision of cardiovascular disease risk estimates to healthcare professionals and patients: a systematic review. *BMJ Open* 2015;5(10):e008717.
21. Karmali KN, Persell SD, Perel P, Lloyd-Jones DM, Berendsen MA, Huffman MD. Risk scoring for the primary prevention of cardiovascular disease. *Cochrane Database Syst Rev* 2017;3:Cd006887.
22. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162(1):55-63.
23. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73.
24. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11(10):e1001744.
25. Wolff R, Collins GS, Kleijnen J, Mallett S, Reitsma JB, Riley R, et al. PROBAST: a risk of bias tool for prediction modelling studies. *24th Cochrane Colloquium*. Seoul, South Korea: Cochrane Database of Systematic Reviews, 2016.
26. Grimshaw JM, Eccles MP, Lavis JN, Hill SJ, Squires JE. Knowledge translation of research findings. *Implement Sci* 2012;7:50.
27. Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gulmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. *Lancet* 2014;383(9912):156-65.

28. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 2014;383(9912):166-75.
29. Chan AW, Song F, Vickers A, Jefferson T, Dickersin K, Gotzsche PC, et al. Increasing value and reducing waste: addressing inaccessible research. *Lancet* 2014;383(9913):257-66.
30. White IR, Rapsomaniki E. Covariate-adjusted measures of discrimination for survival data. *Biom J* 2015;57(4):592-613.
31. Steyerberg E. *Clinical prediction models: a practical approach to development, validation, and updating*: Springer Science & Business Media, 2008.
32. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;35(29):1925-31.

Summary

Prediction models, diagnostic and prognostic, are becoming increasingly important in clinical practice. Unfortunately, research on prediction models is not reproducible and the usefulness of most models in clinical practice is unclear. This is because researchers do not always use the recommended methods for developing or validating a prediction model. Furthermore, often numerous models exist for the same target population or condition. Systematic reviews have therefore become important to appraise and summarize the current evidence on existing prediction models in a specific clinical field. Although ample guidance exists for systematic reviews of interventions and diagnostic tests, guidance for systematic reviews and meta-analyses of prediction models is lacking.

In **Chapter 2** we present guidance for systematic review of prediction models and meta-analysis of the predictive performance of prediction models that were validated across different populations. We describe key steps when performing a systematic review, such as formulating a review question, searching for studies, critical appraisal of identified studies, quantitative data extraction and meta-analysis, and investigating sources of heterogeneity. We also provide recommendations for interpreting the results, and essential items for reporting.

Numerous prediction models are available for the prediction of cardiovascular disease (CVD) in the general population. In **Chapter 3** we present the results of a systematic review in which the current state of CVD risk prediction is summarized, following the guidance presented in Chapter 2. We identified an overabundance of prognostic models for CVD risk prediction. Most of these models predicted the risk of coronary heart disease (CHD) or CVD over 10 years and the majority of models consisted of a similar set of core predictors, including age, gender, smoking, diabetes, blood pressure, and blood cholesterol. Substantial heterogeneity in predictor and outcome definitions was observed between models, and important clinical and methodological information, necessary to externally validate the model or even apply it in clinical practice, were often missing. Only one third of the available models was externally validated, and therefore the usefulness in clinical practice of most models remains unclear. We advise that future research should focus on externally validating and comparing existing models, on tailoring these models to local settings, and investigating whether these models can be extended by addition of new predictors.

In Chapter 3 we noticed that most researchers that externally validated existing prognostic models, focused on the Framingham models. In **Chapter 4** we identified all external validation studies of three often advocated prognostic models (Framingham Wilson, Framingham ATP III, and Pooled Cohort Equations (PCE)) for the prediction of 10-year risk of CHD or CVD, and summarized their predictive performance in terms of discrimination and calibration. There was considerable heterogeneity in the predictive performance between studies, likely due to differences in eligibility criteria, and

population characteristics. On average, however, all models discriminate comparable well and all models overestimate the 10-year risk of CHD and CVD. Overestimation was most pronounced in high-risk individuals and European populations. We highly recommend that researchers further explore reasons for overprediction and that the models be updated for specific populations before using them in clinical practice.

One way to improve the predictive performance of available prediction models, is to add new predictors to the model. Predictor finding studies have reported the association between several biomarkers and the occurrence of CVD. However, in **Chapter 5** we show that adding these extra biomarkers to a prognostic model with traditional predictors did result in very limited improvement of performance of this model for predicting 10-year risk of CHD. Traditional risk factors, like age, smoking, diabetes, blood pressure, and blood cholesterol still seem to be the most important predictors. An alternative strategy to improve predictive performance, which might be much more effective, is to tailor existing prediction models based on traditional predictors only to specific settings using model updating strategies

Chapter 6 focuses on one of the possible explanations suggested to cause overestimation of existing prognostic models for CVD, namely the use of treatment that lowers CVD risk (e.g. antihypertensive or lipid lowering medication) in participants included in studies developing and validating prognostic models. Most studies did not consider the use of any treatment, and even did not describe information on the use of treatment at baseline or during follow-up. Future prognostic model studies should clearly report the use of treatments by study participants and consider the potential impact of treatment use on the study findings.

The lack of crucial information due to poor reporting was not only found in the field of CVD risk prediction, but also in a general set of studies reporting on the development or external validation of diagnostic and prognostic models (**Chapter 7**). More than half of the items that are considered essential for transparent reporting of a prediction model in the TRIPOD statement were not or inadequately reported. We thus concluded that reporting should be improved, by making use of the TRIPOD reporting guideline.

In **Chapter 8** we studied sources of heterogeneity in the predictive performance of prognostic models from various clinical fields. Using a meta-epidemiological approach, we found that this heterogeneity is mainly associated with variation in population aspects and noticed some indications for an association with predictor and outcome measurement methods. Further research is needed to evaluate under what circumstances certain design issues lead to bias in the predictive performance of prognostic models.

In conclusion, we have developed methods for systematic review and meta-analysis of prediction models, thereby encountering several challenges in the design, methodological conduct, reporting, and research focus of primary prognostic model studies. We believe this needs to change, and believe that education is key for properly implementing new methodological and reporting standards.

Based on the results of this thesis, we plan to provide better education to (bio)medical students, develop freely available online education materials for researchers, journal editors, peer reviewers, healthcare providers, and students, raise awareness of the available tools and methods at scientific conferences and publish further guidance and training. Furthermore, we plan to implement the methods we developed for systematic review of prediction model studies within Cochrane.

Samenvatting

Predictiemodellen (voorspellingsmodellen, risicoscores), zowel diagnostische als prognostische, worden steeds belangrijker in de klinische praktijk. Helaas is onderzoek naar predictiemodellen vaak niet reproduceerbaar en is het nut van de meeste modellen in de klinische praktijk onduidelijk. Dit komt doordat onderzoekers niet altijd de aanbevolen methoden gebruiken voor het ontwikkelen of valideren van een predictiemodel. Bovendien bestaan er vaak talloze modellen voor dezelfde subpopulatie of ziekte. Om het huidige bewijsmateriaal over bestaande predictiemodellenmodellen in een specifiek klinisch gebied te beoordelen en samen te vatten, zijn systematische literatuuroverzichten belangrijk. Hoewel er voldoende methodologische kennis bestaat over deze systematische literatuuroverzichten van interventiestudies en diagnostische test studies, ontbreekt een leidraad voor systematische literatuuroverzichten van predictiemodellen.

In **Hoofdstuk 2** presenteren we richtlijnen voor het systematisch samenvatten van literatuur met betrekking tot predictiemodellen en het meta-analyseren van de voorspellende prestaties van modellen die gevalideerd zijn in verschillende populaties. Daarbij beschrijven we de belangrijkste stappen bij het uitvoeren van een systematisch literatuuroverzicht, zoals het formuleren van een onderzoeksvraag, het zoeken naar studies, het kritisch beoordelen van geïdentificeerde studies, kwantitatieve data-extractie en meta-analyse en het onderzoeken van oorzaken van heterogeniteit. We doen ook aanbevelingen voor het interpreteren van de resultaten en rapporteren van essentiële items.

Er zijn tal van predictiemodellen beschikbaar voor het voorspellen van hart- en vaatziekten in de algemene bevolking. In **Hoofdstuk 3** beschrijven we de resultaten van een systematisch literatuuroverzicht waarin we de huidige status van cardiovasculaire risicovoorspelling samenvatten volgens de methodologische stappen beschreven in hoofdstuk 2. We vonden een overvloed aan prognostische modellen voor cardiovasculaire risicovoorspelling. De meeste van deze modellen voorspelden het risico op coronaire hartziekte of hart- en vaatziekte gedurende 10 jaar en de meerderheid van de modellen bestond uit een vergelijkbare set predictoren (voorspellers), waaronder leeftijd, geslacht, roken, diabetes, bloeddruk en cholesterolwaardes in het bloed. De modellen varieerden aanzienlijk in predictor- en uitkomstdefinities. Methodologische informatie die onmisbaar is om een model extern te valideren of toe te passen in de klinische praktijk, ontbrak vaak. Slechts een derde van de beschikbare modellen was extern gevalideerd en daarom blijft het nut van de meeste modellen voor de klinische praktijk onduidelijk. We adviseren dat toekomstig onderzoek zich richt op het extern valideren en vergelijken van bestaande modellen, op het aanpassen van deze modellen op specifieke omstandigheden en het onderzoeken of deze modellen kunnen worden verbeterd door de toevoeging van nieuwe predictoren.

In Hoofdstuk 3 ontdekten we dat de meeste onderzoekers die bestaande prognostische modellen extern valideerden, zich richtten op de Framingham modellen. In **Hoofdstuk 4** hebben we alle externe validatiestudies van drie vaak gebruikte en geëvalueerde Framingham modellen (Framingham Wilson, Framingham ATP III en Pooled Cohort Equations (PCE)) voor de voorspelling van het 10-jarige risico op coronaire hartziekte of hart- en vaatziekte bekeken en hun prestaties (in termen van discriminatie en calibratie) vergeleken. Er waren aanzienlijke verschillen tussen de studies wat betreft de prestaties van de modellen, waarschijnlijk als gevolg van verschillen in de gehanteerde in- en exclusie criteria en kenmerken van de bestudeerde populaties. Gemiddeld echter discrimineerden alle modellen in dezelfde mate tussen mensen mét hart- en vaatziekte en mensen zonder hart- en vaatziekte. Ook overschatten alle modellen het 10-jarige risico op hart- en vaatziekte. Deze overschatting was het grootst in Europese populaties en hoog-risico populaties. We bevelen aan dat onderzoekers de redenen voor deze overschatting verder onderzoeken en dat de modellen eerst worden afgestemd op specifieke populaties, voordat ze in de klinische praktijk worden gebruikt.

Een manier om de prestaties van beschikbare predictiemodellen te verbeteren, is door nieuwe predictoren aan het model toe te voegen. Eerdere studies hebben aangetoond dat er associaties zijn tussen bepaalde biomarkers en het optreden van hart- en vaatziekten. In **Hoofdstuk 5** laten we echter zien dat het toevoegen van deze biomarkers aan een prognostisch model met traditionele predictoren resulteerde in een zeer beperkte verbetering van de prestaties van dit model voor het voorspellen van het 10-jaars risico op coronaire hartziekten. Traditionele predictoren, zoals leeftijd, roken, diabetes, bloeddruk en cholesterol in het bloed, lijken dus nog steeds de belangrijkste predictoren te zijn. Een alternatieve (en waarschijnlijk effectievere) strategie om de prestaties van een model te verbeteren, is om bestaande predictiemodellen (met veelal traditionele predictoren) af te stemmen op specifieke populaties met behulp van zogenaamde 'updating' strategieën.

Hoofdstuk 6 richt zich op één van de mogelijke oorzaken van de overschatting van voorspelde risico's door bestaande prognostische modellen voor hart- en vaatziekten. Mensen die deel uitmaken van studies waarin prognostische modellen worden ontwikkeld en gevalideerd, ondergaan vaak behandelingen die het risico op hart- en vaatziekten verlagen (bijvoorbeeld bloeddruk- of cholesterolverlagende medicatie). In de meeste studies werd het gebruik van deze behandeling niet meegenomen in de analyses en vaak werd er zelfs geen informatie gegeven over het gebruik van de behandeling bij aanvang of tijdens de follow-up van een studie. Toekomstige studies waarin prognostische modellen worden ontwikkeld of gevalideerd, zouden het gebruik van medicatie door studiedeelnemers moeten rapporteren en de potentiële impact van het medicijngebruik op de onderzoeksresultaten in overweging moeten nemen.

Het gebrek aan cruciale informatie vanwege onvolledige rapportage werd niet alleen gevonden op het gebied van cardiovasculaire risicovoorspelling, maar ook in een algemene reeks publicaties betreffende de ontwikkeling of externe validatie van

diagnostische en prognostische modellen (**Hoofdstuk 7**). Meer dan de helft van de items die essentieel worden geacht voor een transparante en volledige rapportage van een predictiemodel werden niet of onvoldoende gerapporteerd. We concludeerden dat de rapportage van dit type studies verbeterd zou moeten worden. Gebruik van de TRIPOD rapportagerichtlijn zou hierbij kunnen helpen.

In **Hoofdstuk 8** hebben we bronnen van heterogeniteit in de prestaties van prognostische modellen uit verschillende klinische gebieden bestudeerd. Gebruikmakend van een meta-epidemiologische benadering, vonden we dat deze heterogeniteit met name geassocieerd is met variatie in populatieaspecten. Ook vonden we enkele aanwijzingen voor een associatie met meetmethoden van predictoren en uitkomsten. Verder onderzoek is nodig om te evalueren onder welke omstandigheden bepaalde methodologische studiekenmerken leiden tot vertekening in de prestaties van prognostische modellen.

Concluderend hebben we methoden ontwikkeld voor systematische literatuuroverzichten en meta-analyses van predictiemodellen, waarbij we verschillende uitdagingen tegenkwamen in de onderzoeksopzet, methodologische kwaliteit en rapportage van primaire prognostische studies. We zijn van mening dat dit moet veranderen en geloven dat onderwijs de sleutel is naar het correct implementeren van nieuwe methodologische en rapportagestandaarden.

Op basis van de resultaten van dit proefschrift zijn we van plan om beter onderwijs te bieden aan (bio) medische studenten en vrij beschikbaar online onderwijsmateriaal te ontwikkelen voor onderzoekers, redacteuren, referenten, zorgverleners en studenten. We willen mensen bewust maken van de beschikbare hulpmiddelen en methoden door deze te presenteren op wetenschappelijke conferenties en verdere begeleiding en training aan te bieden. Bovendien zijn we van plan de door ons ontwikkelde methoden voor systematische literatuuroverzichten van studies naar predictiemodellen te implementeren binnen Cochrane.

Dankwoord

Het zit er (bijna) op! En uiteraard heb ik dit niet alleen gedaan maar met steun van vele anderen. Graag wil ik daarom de volgende mensen bedanken.

Geachte prof. Moons, beste Carl, bedankt dat ik zo veel van je heb mogen leren. Jouw enthousiasme werkt aanstekelijk. Of het nu gaat om het geven van een college, het bedenken van een nieuw onderzoeksplan, of het 'verkopen' van mijn resultaten, overal ga jij vol energie in, en dat werkt erg inspirerend. Ik heb van jou geleerd dat het ook mogelijk is om té kritisch te zijn en dat ik best wat meer plezier mag uitstralen tijdens overleggen en dergelijke. Want ja, ik beleef heel veel plezier aan het werken in dit team! Ik ben jou en Lotty heel dankbaar dat ik de kans krijg om het werk dat ik in het kader van mijn PhD heb gedaan nu te implementeren binnen Cochrane en daarmee de cirkel rond te maken.

Geachte prof. Scholten, beste Rob, gedurende de jaren raakte jij steeds meer betrokken bij mijn onderzoeksprojecten. Omdat ik natuurlijk steeds dieper in de wereld van de predictiemodellen belande was het heel fijn dat jij er was om van een afstandje naar mijn stukken te kijken en me te wijzen op dingen die beter uitgelegd moesten worden. Ook heb ik heel veel respect voor de manier waarop jij omgaat met het ondersteunend personeel, je toegankelijke en open houding is een voorbeeld voor mij.

Geachte Dr. Hooft, lieve Lotty, dankjewel dat je vanaf het allereerste begin vertrouwen in mij had. Dat zorgde ervoor dat ik ook vertrouwen kreeg in mezelf en in mijn kunnen. Je leerde me niet alleen hoe ik het beste een paper moet schrijven, een poster moet maken, of een presentatie moet geven, maar ook om goed voor mezelf te zorgen. We zijn een heel hecht team samen met de andere (affiliated) Cochraners en daar heb jij een heel groot aandeel in.

Geachte Dr. Debray, beste Thomas, in het begin vond ik het soms best lastig om volledig te begrijpen wat je probeerde uit te leggen, maar inmiddels spreken we dezelfde taal (op dat Vlaamse accent na dan). Je nam me mee naar de wereld van standard errors, logit transformaties, en random effects meta-analyse, zaken waar mijn nerdy kant heel blij van wordt! Heel leuk ook dat ik soms bij je thuis werd uitgenodigd voor etentjes.

Dear prof. Visseren, prof. Bots, prof. Collins, prof. Roes, and dr.ir. den Ruijter, thank you for being part of my Assessment Committee. Furthermore, I thank you and dr. van den Bruel for being part of the Doctoral Examination Committee.

Beste Hans, hoewel je officieel geen onderdeel uitmaakt van mijn promotieteam, heb ik je toch altijd beschouwd als onderdeel van mijn begeleidingsteam. Jouw goede inzichten hebben bijgedragen aan de kwaliteit van elk hoofdstuk in dit proefschrift. Wat ik ook heel erg waardeer is jouw inzet voor het hele Julius Centrum. Zeer zeker ben jij de senior met het hoogste opkomstpercentage bij de Julius seminars en methods meetings.

Beste Linda, dankjewel voor de gezellige 'biomarkers-meetings'. Je hebt me niet alleen veel geleerd over Prentice-weighted Cox Proportional Hazards modellen, maar ook over hoe om te gaan met een hoge werkdruk. Hier ben ik je heel dankbaar voor.

Beste René, mijn woensdag begint altijd goed met een gesprekje met jou! Dankjewel dat je me vanalles hebt geleerd over het zoeken naar literatuur, maar ook voor je vragen die me triggerden mijn onderzoek in een breder perspectief te plaatsen.

Lieve Pauline, niet lang nadat we allebei hier in Utrecht begonnen startte ons eerste project samen: TRIPOD adherence. Wat mij betreft een gouden combinatie: we lijken heel erg op elkaar, maar vullen elkaar ook heel mooi aan. De vele koffieautomaatgesprekken (die soms stiekem best iets té lang duurden) waren altijd een fijne onderbreking van de dag. Toen bleek dat we allebei dezelfde groepsreis in Zuid-Afrika wilde gaan doen heb ik ook geen moment getwijfeld: als er een college is waarmee ik op vakantie zou kunnen/willen, dan ben jij dat wel! Het is een geruststellende gedachte dat jij op 14 juni achter mij staat.

Dear Romin, hours and hours we spend discussing the selection of articles, data extraction, and risk of bias in one of the windowless meeting rooms. I really enjoyed these meetings and learned a lot from them! I specifically value your interest in the work I'm doing and that you always take time to comment on papers or abstracts.

Beste Michiel, je bent er nog niet zo lang, en je bent er ook niet zo vaak, maar we hebben al veel gezellige koffie-momentjes gehad. Ik vind het leuk dat je altijd interesse toont in mijn werkzaamheden, en hoop dat je nog lang deel zal uitmaken van Cochrane Nederland.

Special thanks to all co-authors of the papers that are included in this thesis. Your input and feedback greatly improved the content of the papers.

(Ex)-leden van het methodologie team, in het speciaal Christiana, Ewoud, Josan, Kevin, Lisette, Maarten, Rolf, Sander, Sjoerd, Valentijn. Bedankt voor de gezelligheid en voor jullie waardevolle input bij mijn presentaties. Het is altijd fijn als iemand met een frisse blik kijkt naar je bevindingen.

Thank you dearest roommates of room 6.118, Amy, Elsbeth, Femke, Fien, Giske, Josefien, Linda, Ly, Mansour, Marian, Marit, Meander, and Tom. I really liked our lunches, walks, dinners and other activities. Femke, Giske en Linda, we begonnen ongeveer tegelijk op kamer 6.118 en gingen samen de master doen. De afgelopen jaren zijn jullie alle drie moeder geworden. Ik vond het geweldig om dit mee te mogen maken. Nu ik nog hè! ;-)

Lieve oud-huisgenootjes van gang 40, Dimitrij, Josianne, Liset, Marlies, Robin en Yvette. Ik ben blij dat we nog steeds contact hebben met elkaar. Jarenlang (inclusief de eerste maanden dat ik in Utrecht werkte) hebben we lief en leed met elkaar gedeeld, urenlang in de keuken (of op de gang) gekletst, en veel plezier gemaakt.

Lieve SQL half 5-ers, Chris, Guus, Guido, Janneke, Joris, Lieke, Marcel, Marga, Marlies en Toon. In september is het 10 jaar geleden dat we elkaar voor het eerst ontmoetten. Sindsdien hebben we vele weekendjes weg, activiteiten, etentjes en avondjes uit gehad. Het allerfijnste aan deze groep vind ik dat iedereen zichzelf kan en mag zijn. Lieke, ik vind het super leuk dat jij als paranimf achter mij staat op deze belangrijke dag!

Lieve schoonfamilie, Leo, José, Rob, Tom & Kristel, oma Hanssen, JES'ers, en familie Hanssen, fijn dat jullie mij zo hebben opgenomen in jullie familie. Ik heb me altijd erg welkom gevoeld. De familieweekenden en -dagen konden soms best vermoeiend zijn (al die honden ook!!), maar waren ook altijd heel gezellig.

Lieve Kenny en Henny, ondanks dat we elkaar veel te weinig zien, voel ik een enorm sterke band met jullie. Wat ben ik blij dat jij mijn peetoom bent, Kenny.

Lieve papa, mama, Toine & Christel, Nikkie, oma Nel en oma Diny, jullie vormen mijn veilige haven waar ik altijd op kan terugvallen. Bedankt voor jullie steun, zelfs als ik het veel te druk had om bij jullie langs te komen waren jullie er voor mij.

Lieve Anouk, dankjewel dat je altijd naast mij staat. Bij jou kan ik volledig mezelf zijn.

Curriculum Vitae

Anneke Damen was born on October 10, 1990 in Oosterhout, the Netherlands. In 2008, she graduated from the Mgr. Frencken college in Oosterhout and started with a bachelor Biomedical Sciences at Radboud University Nijmegen. During her master Biomedical Sciences, she specialized in epidemiology and pathobiology. As part of her bachelor and master, Anneke did several internships on various topics at the Radboudumc department of Health Evidence, the National Expert and Training Centre for Breast Cancer Screening, the Radboudumc department of Pathology, and at Sanquin blood supply, under the supervision of Dr. Mireille Broeders, Dr. Janine Timmers, Dr. William Leenders, and Dr. Pieternel Pasker-de Jong.

In 2014, Anneke started working as a PhD student at the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, supervised by Prof. Carl Moons, Prof. Rob Scholten, Dr. Lotty Hooft, and Dr. Thomas Debray. Her project focused on systematic reviews and meta-analyses of prediction model studies. She combined her PhD project with the postgraduate master Clinical Epidemiology. Furthermore, she worked as affiliated researcher for Cochrane Netherlands for which she was active as a teacher on several courses and worked on projects for the Belgian Health Care Knowledge Centre and Zorginstituut Nederland.

Anneke currently works as a postdoctoral researcher for Cochrane Netherlands and the Julius Center for Health Sciences and Primary Care, and is the coordinator of the Cochrane Prognosis Methods Group. She is project leader of the implementation of reviews of prognosis studies within Cochrane.