Prognostic research in treated populations

Romin Pajouheshnia

Prognostic research in treated populations.

PhD thesis. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

Author Romin Pajouheshnia

The studies in this thesis were funded by the Netherlands Organization for Scientific Research (projects 9120.8004 and 918.10.615). Financial support by the Julius Center for Health Sciences and Primary Care for the publication of this thesis is gratefully acknowledged.

ISBN 978-94-93019-95-9
Design ProefschriftMaken | proefschriftmaken.nl
Printed by ProefschriftMaken | proefschriftmaken.nl

Prognostic research in treated populations

Prognostische onderzoek met behandelde populaties
(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 colege voor promoties
in het openbaar te verdedigen op

dinsdag 30 oktober 2018
des middags te 4:15 uur

door
Romin Pajouheshnia
geboren op 21 oktober 1988
te Plymouth, United Kingdom

Promotor:	Prof. dr. K.G.M. Moons
Copromotoren:	Dr. L.M. Peelen Dr. R.H.H. Groenwold

Contents

CHAPTER 1

General introduction

What are clinical prediction models?

Prediction models are an important tool in modern, evidence-based clinical practice.[1-5] Prediction models can be broadly divided into two categories, either diagnostic- to predict the presence of a condition or health status or prognostic- to predict the future occurrence of a health outcome or a health trajectory.[6, 7] In both cases, the models combine and weight clinical and demographic information from an individual to commonly provide an estimated probability of having or developing a certain outcome.[8, 9] The predicted probabilities provided by such models can be used to guide or support the decisions made by health professionals and patients. Several studies have demonstrated the added value of using model-based predictions over clinical gestalt, with gains in accuracy, efficiency, consistency and cost-effectiveness frequently cited as a potential benefits.[10-12]

The evidence basis for prediction models is founded on the use of (large) data sets from relevant populations and settings for their development and validation.[13] The study populations in prediction model studies should be representative of individuals for whom predictions will be made in practice- the "target population" for the prediction model. Once a model has been developed, the model must be externally validated, i.e. tested or evaluated in terms of its predictive accuracy,[14-17] before being assessed in terms of its impact on clinical decisions and the health of individuals, and finally implemented in routine practice.

What do we expect from a good prediction model?

When validating a prediction model, there are a number of metrics that are commonly considered.[4, 18] A summary of these measures is presented in Table 1. In essence, predictions made by the model should be as accurate as possible, in order to improve decision making or patient counselling.

Crucially, a prediction model should not only perform well for the study participants from whose data the model was derived, but also in other individuals, usually from the model's target population. These individuals may come from a different geographical location or a different health setting with a different standard of care (e.g. primary care vs. secondary care),[19] and thus may come from a population with a substantially different distribution of risk factors.[16, 20, 21] As a result, the goal of developing prediction models that can be generalized to future individuals remains challenging.

Table 1: A summary of common measures of prediction model performance.

Performance criteria	Explanation	Common metrics
Discrimination	How well a prediction model can correctly separate cases (do/will have the outcome) from non-cases (do not /will not have the outcome).	• Area under the ROC curve (AUC) or c-statistic • D-statistic • Discrimination slope
Calibration	The agreement between the predicted probabilities provided by the model and actual outcomes of individuals.	• Calibration intercept, slope • Observed:expected ratio
Overall performance	The overall accuracy of predictions or degree to which the model fits the data.	• R^2 • Brier score
Clinical utility	The benefit of using a prediction model in decision making.	• Decision curve • Net benefit

The problem: heterogeneous and sub-optimal model performance

In recent years, a number of systematic reviews of prediction models have been published, across a range of medical areas. Common findings of these studies include poorer performance of prediction models when validated in external data sets and considerable heterogeneity in the performance of prediction models from study to study.[22, 23] Clearly, poor performance across populations in terms of the measures listed in Table 1 can seriously limit the usefulness of a prediction model. Similarly, heterogeneity in the performance of a prediction model poses a substantial challenge for end-users (e.g. health care professionals and policy makers) when trying to determine whether or not a prediction model is suitable for use. As a result, when considering a particular prediction model, end-users are left struggling with the following questions:

- Will the model provide accurate predictions for future individuals?
- For which patients is the prediction model most suitable?
- Will the model transport well to different settings?
- If there are competing models, which model should be preferred for a specific patient?

Heterogeneity in the performance of prediction models is commonly attributed to differences in the associations between predictors and outcomes across populations, and variation in participant characteristics or "case-mix" in the presence of interactions between these characteristics and unmeasured variables, a phenomenon that has been termed the "spectrum effect".[24] As a consequence, recommendations have been made for the interpretation of the findings of prediction model performance in the presence of case-mix variation across model validation studies.[25-29] However, there is a growing body of evidence that additional factors influence prediction model performance when evaluated across different study populations.[23, 30-33] An issue of primary concern is bias in

the development and validation of prediction models due to mismatches between what studies intend to estimate, and the methods and data used in the studies.

Treatment use in prediction studies

Typically, the data used to develop or validate clinical prediction models come from individuals who receive treatment (e.g. medication or surgical procedures) or health interventions (e.g. changes in care or monitoring, or lifestyle changes), especially in the context of disease prevention. Treatment can be considered as one aspect of the case-mix of a population, and can vary greatly depending from setting to setting, depending on a number of factors such as the policy or protocols within a setting, cultural preferences or trends (over time), or the availability of certain interventions. On top of this, concerns have been raised over the use of treatments in prognostic prediction studies when interest lies in making predictions of the risk of certain future outcomes if treatment were withheld- to estimate "natural prognosis", for instance.[34-38] In this case, there is an urgent need for a greater understanding of how such treatments, alongside other poorly understood factors such as differences in the way predictors are measured across populations and differences in the distribution and effects of predictors across populations in the presence of interactions with unmeasured variables, could affect the performance of prognostic models

Outline of this thesis

This thesis aims to develop and improve methodology for prediction model research, with a focus on prognostic model studies, and in turn address a number of ongoing research questions:

- When does treatment use in a prognostic study lead to bias?
- What additional factors can cause prediction model performance to vary across populations?
- When and how should treatment use and other factors that affect prediction model performance be taken into account when developing or validating a prediction model?

Chapter 2 provides an example of the heterogeneity seen in prognostic model performance across populations. A systematic review and meta-analysis of the predictive performance of three cardiovascular risk prediction models is presented. Risk of bias is assessed and meta-regression analysis is used to identify sources of heterogeneity in model performance between the studies.

Chapters 3, 4 and 5 address the challenge that treatment use creates for prediction-particularly prognostic- model research. In chapter 3, a typology for treatments used in prognostic model studies is proposed, in order to better understand when and how they should be taken into account. Existing practices when handling treatment use in prognostic research are identified in a systematic review of cardiovascular prediction model studies. Chapter 4 specifically addresses the impact of the use of treatments when validating a prognostic model. The consequences of ignoring treatment use in a validation sample are explained, and methods to correct for the use of treatments are explored theoretically and tested using simulated data. Chapter 5 compares statistical methods to account for the effects of time-varying treatment use when developing a prognostic model using observational data from patients who received beta-blocker treatment during follow-up.

In chapter 6, a second source of heterogeneity in prediction model performance is investigated, namely, differences in the measurement of predictors across settings. The effect of variation in how a predictor is measured from sample to sample (considered formally in terms of "measurement errors") on the AUC of a prediction model is described analytically and examined in a case study of a diagnostic prediction model.

Chapter 7 builds on the findings from previous chapters to provide recommendations for "best practices" when developing or validating prognostic models using data from randomized trials. Special attention is given to factors that could compromise the validity and generalizability of prognostic model research conducted using data from RCTs.

Finally, this thesis ends with the proposal of a framework of "prediction estimands", with the dual aim of improving the interpretation of prediction models and the heterogeneity in their performance, as well as the quality of future prediction model research.

References

1. Adams ST, Leveson SH. Clinical prediction rules. BMJ. 2012;344.
2. Beattie P, Nelson R. Clinical prediction rules: what are they and what do they tell us? Australian Journal of Physiotherapy. 2006;52(3):157-63.
3. 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.
4. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating: Springer Science & Business Media; 2008.
5. Plüddemann A, Wallace E, Bankhead C, Keogh C, Van der Windt D, Lasserson D, et al. Clinical prediction rules in practice: review of clinical guidelines and survey of GPs. The British Journal of General Practice. 2014;64(621):e233-e42.
6. Grobbee DE, Hoes AW. Clinical epidemiology: Jones & Bartlett Publishers; 2014.
7. 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.
8. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. Jama. 1997;277(6):488-94.
9. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? Bmj. 2009;338:b375.
10. Williams C, Brunskill S, Altman D, Briggs A, Campbell H, Clarke M, et al. Cost-effectiveness of using prognostic information to select women with breast cancer for adjuvant systemic therapy. 2006.
11. Sanders SL, Rathbone J, Bell KJL, Glasziou PP, Doust JA. Systematic review of the effects of care provided with and without diagnostic clinical prediction rules. Diagnostic and Prognostic Research. 2017;1(1):13.
12. Johnston ME, Langton KB, Haynes RB, Mathieu A. Effects of computer-based clinical decision support systems on clinician performance and patient outcome. A critical appraisal of research. Ann Intern Med. 1994;120(2):135-42.
13. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLOS Medicine. 2013;10(2):e1001381.
14. Bleeker S, Moll H, Steyerberg E, Donders A, Derksen-Lubsen G, Grobbee D, et al. External validation is necessary in prediction research:: A clinical example. Journal of clinical epidemiology. 2003;56(9):826-32.
15. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245-7.
16. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. Bmj. 2009;338:b605.
17. 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.

18. 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.
19. Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: A review. Journal of Clinical Epidemiology. 2008;61(11):1085-94.
20. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19(4):453-73.
21. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Annals of internal medicine. 1999;130(6):515-24.
22. Haskins R, Cook C. Enthusiasm for prescriptive clinical prediction rules (eg, back pain and more): a quick word of caution. British Journal of Sports Medicine. 2016.
23. 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.
24. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. BMJ. 2016;353.
25. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. J Clin Epidemiol. 2016;79:76-85.
26. Vergouwe Y, Moons KGM, Steyerberg EW. External Validity of Risk Models: Use of Benchmark Values to Disentangle a Case-Mix Effect From Incorrect Coefficients. American Journal of Epidemiology. 2010;172(8):971-80.
27. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, 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.
28. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. Journal of Clinical Epidemiology. 2015;68(3):279-89.
29. Nieboer D, van der Ploeg T, Steyerberg EW. Assessing Discriminative Performance at External Validation of Clinical Prediction Models. PLoS One. 2016;11(2):e0148820.
30. Ban J-W, 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.
31. Charlson ME, Ales KL, Simon R, MacKenzie C. Why predictive indexes perform less well in validation studies: Is it magic or methods? Archives of Internal Medicine. 1987;147(12):2155-61.
32. Khudyakov P, Gorfine M, Zucker D, Spiegelman D. The Impact of Covariate Measurement Error on Risk Prediction. Statistics in medicine. 2015;34(15):2353-67.
33. van Doorn S, Brakenhoff TB, Moons KGM, Rutten FH, Hoes AW, Groenwold RHH, et al. The effects of misclassification in routine healthcare databases on the accuracy of prognostic prediction models: a case study of the CHA2DS2-VASc score in atrial fibrillation. Diagnostic and Prognostic Research. 2017;1(1):18.

34. Bernat JL. Ethical aspects of determining and communicating prognosis in critical care. Neurocrit Care. 2004;1(1):107-17.
35. Cheong-See F, Allotey J, Marlin N, Mol BW, Schuit E, Riet G, et al. Prediction models in obstetrics: understanding the treatment paradox and potential solutions to the threat it poses. BJOG: An International Journal of Obstetrics & Gynaecology. 2016;123(7):1060-4.
36. Groenwold RHH, Moons KGM, Pajouheshnia R, Altman DG, Collins GS, Debray TPA, et al. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. Journal of Clinical Epidemiology. 2016;78:90-100.
37. Liew SM, Doust J, Glasziou P. Cardiovascular risk scores do not account for the effect of treatment: a review. Heart. 2011;97(9):689.
38. Thornley S. Studies of cardiovascular disease risk estimation: how, and whether, to account for the effect of drug treatment? ResearchSpace@ Auckland; 2014.

CHAPTER 2

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

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 high-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 1).[9-12] However, some reports have 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.

Box 1: Terminology

	Definition
Case-mix	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.[71]
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.[72, 73]
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.[74]
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).

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.

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)[16] 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, 17] 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.[18] 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,[19,20] see Box 1) 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.[21]

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 [16] and a preliminary version of the Cochrane Prediction study Risk Of Bias Assessment Tool (PROBAST).[22, 23] 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), and analyses (e.g. handling of censoring). After several rounds of piloting and adjusting the data extraction form in teams 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-year total OE ratio and the c-statistic. Based on previous recommendations,[15, 24] 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).[25] 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,[26] using the packages metafor,[27] mvmeta,[28] metamisc,[29] and lme4.[30]

Results

Identification and selection of studies

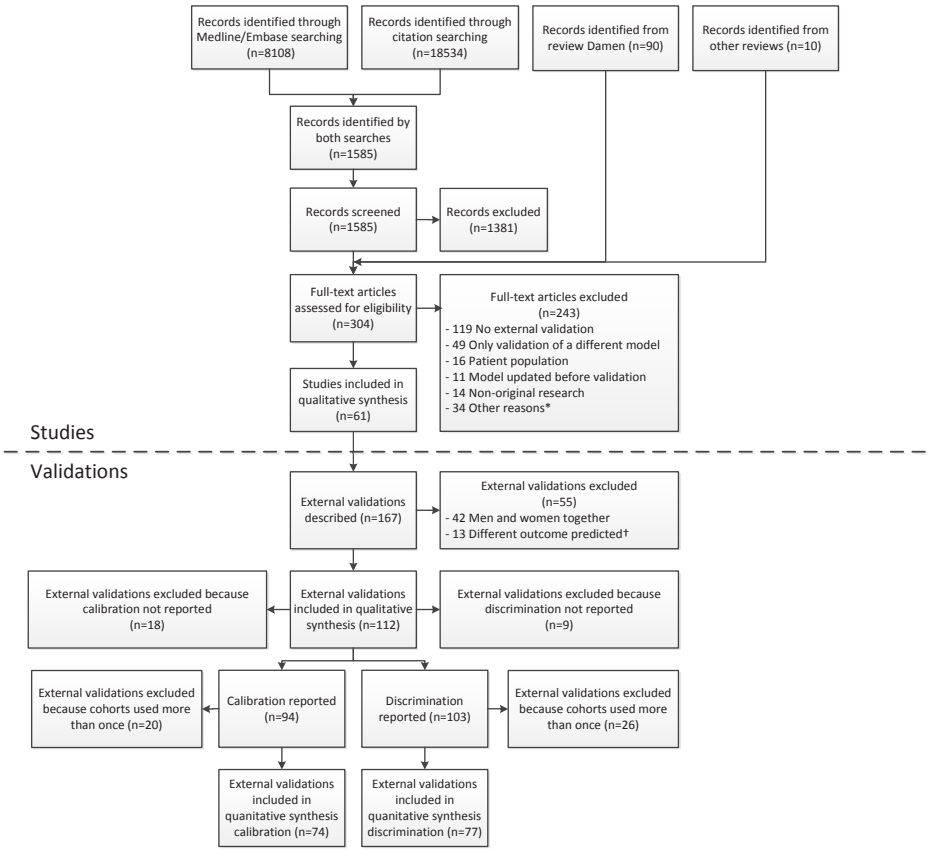
We first identified 100 potentially eligible studies from previously conducted systematic reviews. An additional search identified 1585 studies since June 2013 (Figure 1). Of these 1685 studies, 304 studies were screened on full-text and data were extracted for 61 studies, describing 167 validations 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

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).

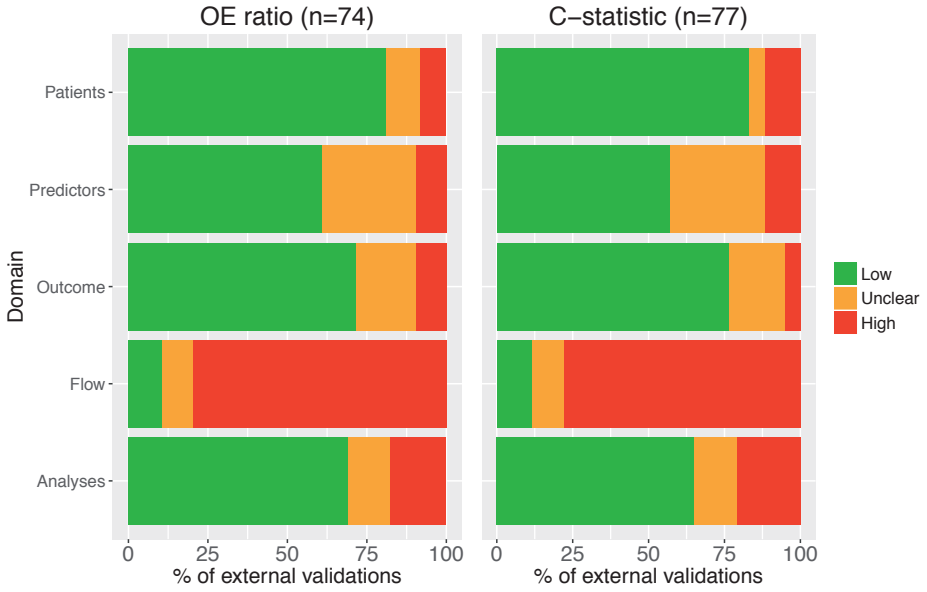


1965 and 2008, and originated from North America (56), Europe (29), Asia (25), and Australia (2). We excluded 18 and 9 external validations because the OE ratio and c-statistic, respectively, were not available, and subsequently excluded 20 and 26 external validations for the OE ratio and c-statistic, 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 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,[31] and one in community-dwelling individuals aged 70–79[32]). For PCE, underestimation of the number of events occurred in Chinese[33] and Korean[34] populations.

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).[35] 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 of around 0.8, suggesting that these models were overfitted or do 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).

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).

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.

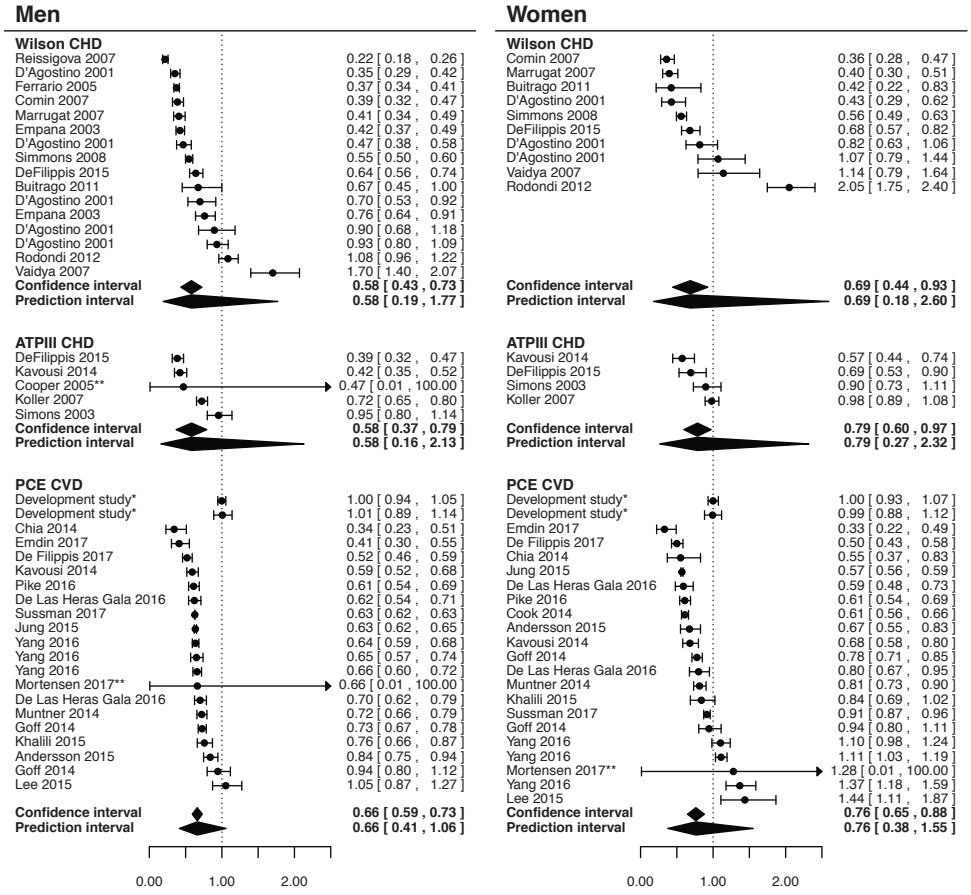


Figure 4: Calibration plots of the Framingham Wilson, ATP III and PCE models. Each line represents one external validation.

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%.[35] CI: confidence interval, PI: prediction interval.

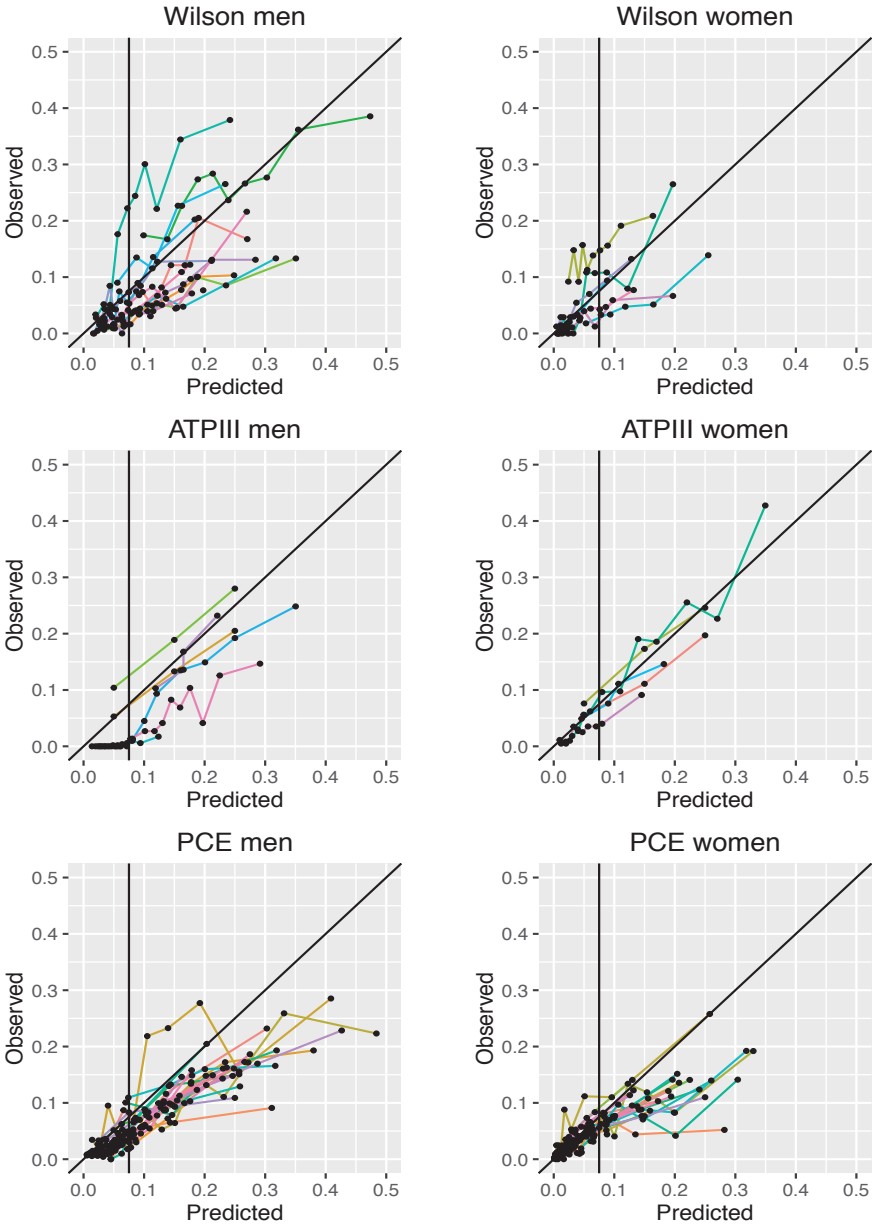
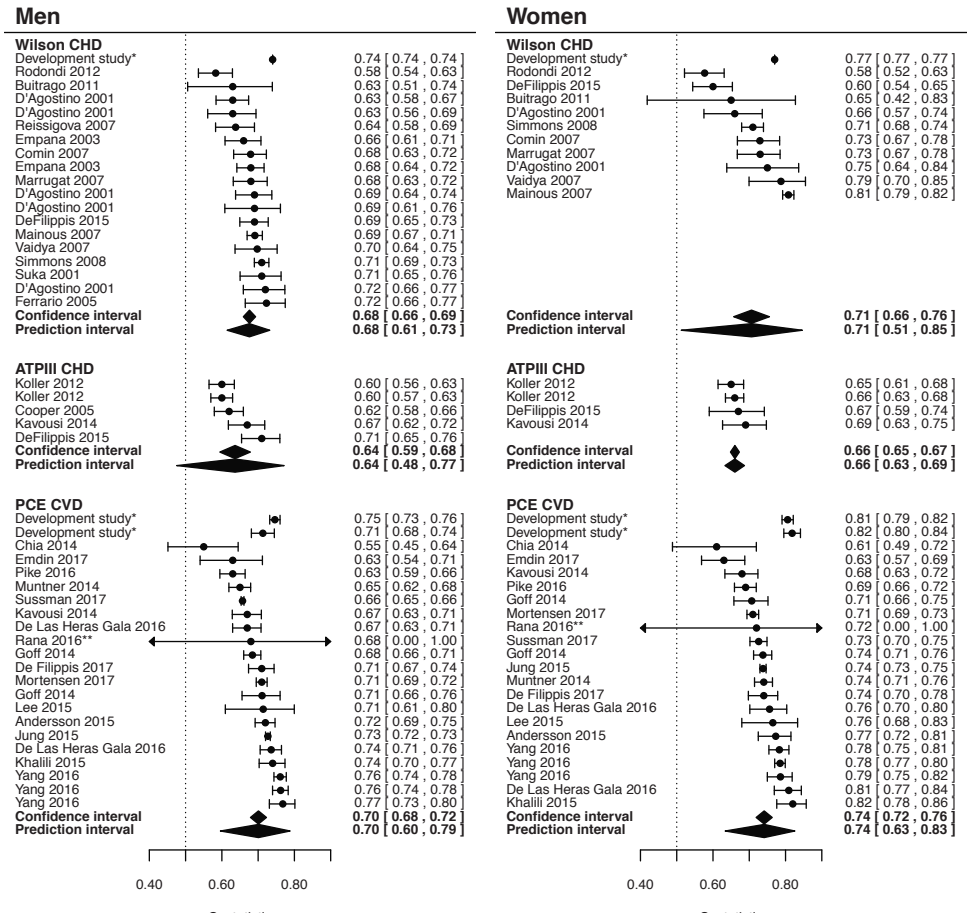


Figure 5: C-statistic in external validations, with 95% confidence intervals and 95% prediction intervals per model.

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.



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.

Discussion

Summary of findings

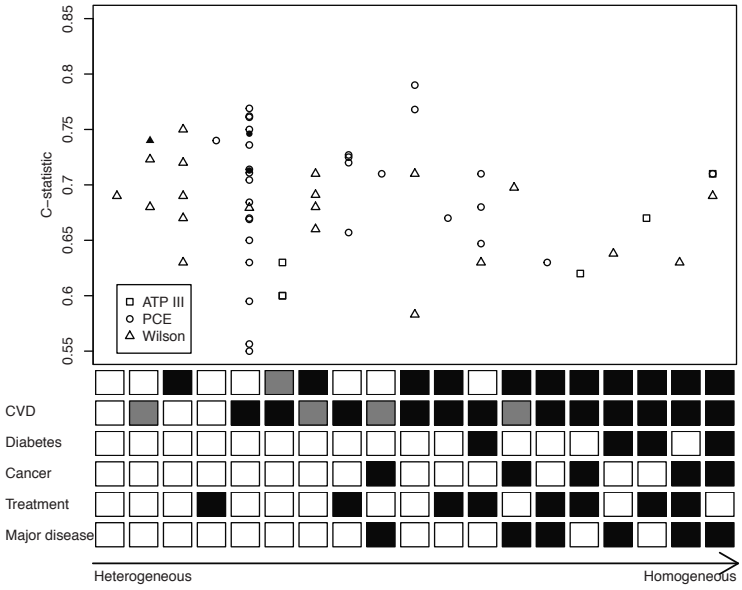
We systematically reviewed the performance of the Framingham Wilson, Framingham ATP III, and PCE models for predicting 10-year risk of CHD or CVD for men and women separately in the general population. We found only small differences in pooled performance between the three models, but large differences 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.[36] In agreement with previous studies,[17, 37-39] 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.[40-42]

Figure 6: C-statistic for different combinations of eligibility criteria.

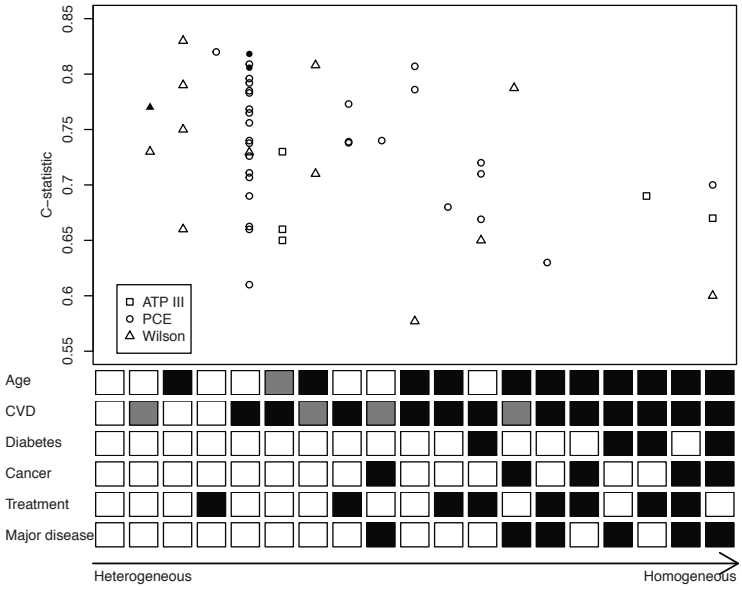
Open squares, circles and triangles represent validations of the ATP III, PCE and Wilson model, respectively. Black circles and triangles represent the studies that developed the PCE models for Whites and African-Americans, and the Wilson model.

Lower panel. Age: white- broad age range included (upper - lower age limit >30 years), black- narrow age range included (upper - lower age limit ≤30 years), grey- age not reported. CVD: white- no exclusion of people with CHD/ CVD, grey- people with previous CHD events were excluded, black- people with previous CVD events were excluded. Diabetes, cancer, major disease: white- no exclusions reported, black- people with these conditions were excluded. Treatment: white- no exclusions, black- people on CVD risk-lowering treatment (e.g. antihypertensives) were excluded.

Men



Women



Reasons for overprediction

There could be several reasons for the observed overprediction, which have also extensively been discussed previously with regards to the PCE.[39, 43, 44] 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.[45] 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.[46-48] 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,[17, 37] 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.[48] 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,[49, 50] and the guidance on adjusting for treatment use in prediction modelling studies,[47, 48, 51] 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.[52-56] Others have however shown that the overestimation could not be fully explained by treatment use and missed outcome events.[46, 57]

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.[58] Hence, one external validation may not be sufficient to claim adequate performance and multiple validations are necessary to get an insight in the generalizability of prediction models.[42] 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, 34, 59, 60] As previously emphasized, more extensive revision methods are often not needed.[19, 20, 61] 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, 62] 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,[63] and the SCORE model which has been tailored to many European countries using national mortality statistics.[64-67] 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.[68] 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[18, 69]), 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[70] and SCORE.[64] 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,[19, 20, 61] 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 and Dr. Zvárová.

Source of funding

Financial support was received from the Cochrane Methods Innovation Funds Round 2 (MTH001F) and Cochrane Trusted methods and Support for Cochrane Reviews of Prognostic studies. KGMM received a grant from the Netherlands Organization for Scientific Research (ZONMW 918.10.615 and 91208004) and from the CREW NHS project, grant number 2013T083, co-funded by the Dutch Heart Society. TPAD was supported by the Netherlands Organization for Scientific Research (91617050).

References

1. WHO. Cardiovascular diseases (CVDs) Fact sheet N°317. 2016.
2. Korczak D, Dietl M and 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 and 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, Ward K and Ebrahim S. 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, Goldberg AC, Gordon D, Levy D, Lloyd-Jones DM, McBride P, Schwartz JS, Shero ST, Smith SC, Jr., Watson K, Wilson PW, Eddleman KM, Jarrett NM, LaBresh K, Nevo L, Wnek J, Anderson JL, Halperin JL, Albert NM, Bozkurt B, Brindis RG, Curtis LH, DeMets D, Hochman JS, Kovacs RJ, Ohman EM, Pressler SJ, Sellke FW, Shen WK, Smith SC, Jr. and Tomaselli GF. 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: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;106:3143-421.
7. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H and Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97:1837-47.
8. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GC, Chiocchia V, Roberts C, Schlussel MM, Gerry S, Black JA, Heus P, van der Schouw YT, Peelen LM and Moons KG. 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, Ikram MA, Stricker BH, Hofman A and Franco OH. 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;311:1416-23.
10. DeFilippis AP, Young R, Carrubba CJ, McEvoy JW, Budoff MJ, Blumenthal RS, Kronmal RA, McClelland RL, Nasir K and Blaha MJ. An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort. *Ann Intern Med*. 2015;162:266-75.
11. Reissigova J and Zvarova J. The Framingham risk function underestimated absolute coronary heart disease risk in Czech men. *Methods Inf Med*. 2007;46:43-9.
12. Comin E, Solanas P, Cabezas C, Subirana I, Ramos R, Gene-Badia J, Cordon F, Grau M, Cabre-Vila JJ and Marrugat J. Estimating cardiovascular risk in Spain using different algorithms. *Rev Esp Cardiol*. 2007;60:693-702.

13. Khalili D, Asgari S, Hadaegh F, Steyerberg EW, Rahimi K, Fahimfar N and Azizi F. 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;184:587-594.
14. D'Agostino RB, Sr., Grundy S, Sullivan LM and Wilson P. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA*. 2001;286:180-7.
15. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, Riley RD and Moons KG. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356:i6460.
16. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, Reitsma JB and Collins GS. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11:e1001744.
17. Beswick AD, Brindle P, Fahey T and 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.
18. Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC, Jr., Sorlie P, Stone NJ, Wilson PW, Jordan HS, Nevo L, Wnek J, Anderson JL, Halperin JL, Albert NM, Bozkurt B, Brindis RG, Curtis LH, DeMets D, Hochman JS, Kovacs RJ, Ohman EM, Pressler SJ, Sellke FW, Shen WK and Tomaselli GF. 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:S49-73.
19. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ and Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med*. 2004;23:2567-86.
20. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE and Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61:76-86.
21. Tramer MR, Reynolds DJ, Moore RA and McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. *BMJ*. 1997;315:635-40.
22. Wolff R, Collins GS, Kleijnen J, Mallett S, Reitsma JB, Riley R, Westwood M, Whiting P and Moons KG. PROBAST: a risk of bias tool for prediction modelling studies. Paper presented at: 24th Cochrane Colloquium; 2016; Seoul, South Korea.
23. Ensor J, Riley RD, Moore D, Snell KI, Bayliss S and Fitzmaurice D. Systematic review of prognostic models for recurrent venous thromboembolism (VTE) post-treatment of first unprovoked VTE. *BMJ open*. 2016;6:e011190.
24. Snell KI, Ensor J, Debray TP, Moons KG and 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.
25. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG and Riley RD. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol*. 2015.

26. R: A language and environment for statistical computing [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2016.
27. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36:1-48.
28. Gasparrini A, Armstrong B and Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Stat Med*. 2012;31:3821-39.
29. Debray TP. Metamisc: Diagnostic and Prognostic Meta-Analysis. 2017.
30. Bates D, Mächler M, Bolker B and Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67.
31. Vaidya D, Yanek LR, Moy TF, Pearson TA, Becker LC and 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;100:1410-5.
32. Rodondi N, Locatelli I, Aujesky D, Butler J, Vittinghoff E, Simonsick E, Satterfield S, Newman AB, Wilson PWF, Pletcher MJ, Bauer DC and Health ABCS. Framingham risk score and alternatives for prediction of coronary heart disease in older adults. *PLoS One*. 2012;7:e34287.
33. Lee CH, Woo YC, Lam JKY, Fong CHY, Cheung BMY, Lam KSL and Tan KCB. Validation of the Pooled Cohort equations in a long-term cohort study of Hong Kong Chinese. *J Clin Lipidol*. 2015;9:640-646.
34. Jung KJ, Jang Y, Oh DJ, Oh BH, Lee SH, Park SW, Seung KB, Kim HK, Yun YD, Choi SH, Sung J, Lee TY, Kim SH, Koh SB, Kim MC, Chang Kim H, Kimm H, Nam C, Park S and Jee SH. The ACC/AHA 2013 pooled cohort equations compared to a Korean Risk Prediction Model for atherosclerotic cardiovascular disease. *Atherosclerosis*. 2015;242:367-375.
35. Eckel RH, Jakicic JM, Ard JD, de Jesus JM, Houston Miller N, Hubbard VS, Lee IM, Lichtenstein AH, Loria CM, Millen BE, Nonas CA, Sacks FM, Smith SC, Jr., Svetkey LP, Wadden TA, Yanovski SZ, Kendall KA, Morgan LC, Trisolini MG, Velasco G, Wnek J, Anderson JL, Halperin JL, Albert NM, Bozkurt B, Brindis RG, Curtis LH, DeMets D, Hochman JS, Kovacs RJ, Ohman EM, Pressler SJ, Sellke FW, Shen WK and Tomaselli GF. 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:S76-99.
36. Paynter NP, Everett BM and Cook NR. Cardiovascular disease risk prediction in women: is there a role for novel biomarkers? *Clin Chem*. 2014;60:88-97.
37. Brindle P, Beswick A, Fahey T and Ebrahim S. Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. *Heart*. 2006;92:1752-9.
38. Eichler K, Puhan MA, Steurer J and Bachmann LM. Prediction of first coronary events with the Framingham score: a systematic review. *Am Heart J*. 2007;153:722-31, 731.e1-8.
39. Cook NR and Ridker PM. Calibration of the Pooled Cohort Equations for Atherosclerotic Cardiovascular Disease: An Update. *Ann Intern Med*. 2016.
40. Knottnerus JA. Diagnostic prediction rules: principles, requirements and pitfalls. *Prim Care*. 1995;22:341-63.
41. Vergouwe Y, Moons KG and Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol*. 2010;172:971-80.

42. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW and Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2014.
43. Cook NR and Ridker PM. Further insight into the cardiovascular risk calculator: the roles of statins, revascularizations, and underascertainment in the Women's Health Study. *JAMA internal medicine*. 2014;174:1964-71.
44. Ridker PM and Cook NR. The Pooled Cohort Equations 3 Years On: Building a Stronger Foundation. *Circulation*. 2016;134:1789-1791.
45. Tobert JA. Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. *Nat Rev Drug Discov*. 2003;2:517-26.
46. Cook NR and 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;174:1964-71.
47. Pajouheshnia R, Peelen LM, Moons KGM, Reitsma JB and Groenwold RHH. Accounting for treatment use when validating a prognostic model: a simulation study. *BMC Med Res Methodol*. 2017;17:103.
48. Pajouheshnia R, Damen JA, Groenwold RH, Moons KG and Peelen LM. Treatment use in prognostic model research: a systematic review of cardiovascular prognostic studies. *Diagnostic and prognostic research*. 2017;1:15.
49. Collins GS, Reitsma JB, Altman DG and Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162:55-63.
50. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF and Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162:W1-73.
51. Groenwold RH, Moons KG, Pajouheshnia R, Altman DG, Collins GS, Debray TP, Reitsma JB, Riley RD and Peelen LM. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *J Clin Epidemiol*. 2016;78:90-100.
52. Muntner P, Safford MM, Cushman M and Howard G. Comment on the reports of over-estimation of ASCVD risk using the 2013 AHA/ACC risk equation. *Circulation*. 2014;129:266-7.
53. Muntner P, Colantonio LD, Cushman M, Goff DC, Jr., Howard G, Howard VJ, Kissela B, Levitan EB, Lloyd-Jones DM and Safford MM. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *JAMA*. 2014;311:1406-15.
54. Krumholz HM. The new cholesterol and blood pressure guidelines: perspective on the path forward. *JAMA*. 2014;311:1403-5.
55. Goff DC, Jr., D'Agostino RB, Sr., Pencina M and Lloyd-Jones DM. Calibration and Discrimination Among Multiple Cardiovascular Risk Scores in a Modern Multiethnic Cohort. *Ann Intern Med*. 2015;163:68.
56. Spence JD. Statins and ischemic stroke. *JAMA*. 2014;312:749-50.

57. Cook NR and Ridker PM. Response to Comment on the reports of over-estimation of ASCVD risk using the 2013 AHA/ACC risk equation. *Circulation*. 2014;129:268-9.
58. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG and Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ*. 2016;353:i3140.
59. Marrugat J, Subirana I, Comin E, Cabezas C, Vila J, Elosua R, Nam BH, Ramos R, Sala J, Solanas P, Cordon F, Gene-Badia J and D'Agostino RB. Validity of an adaptation of the Framingham cardiovascular risk function: the VERIFICA Study. *J Epidemiol Community Health*. 2007;61:40-7.
60. Andersson C, Enserro D, Larson MG, Xanthakis V and 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. *J Am Heart Assoc*. 2015;4.
61. Vergouwe Y, Nieboer D, Oostenbrink R, Debray TP, Murray GD, Kattan MW, Koffijberg H, Moons KG and Steyerberg EW. A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med*. 2016.
62. Grundy SM, D'Agostino Sr RB, Mosca L, Burke GL, Wilson PW, Rader DJ, Cleeman JI, Roccella EJ, Cutler JA and Friedman LM. Cardiovascular risk assessment based on US cohort studies: findings from a National Heart, Lung, and Blood institute workshop. *Circulation*. 2001;104:491-6.
63. Hajifathalian K, Ueda P, Lu Y, Woodward M, Ahmadvand A, Aguilar-Salinas CA, Azizi F, Cifkova R, Di Cesare M, Eriksen L, Farzadfar F, Ikeda N, Khalili D, Khang YH, Lanska V, Leon-Munoz L, Magliano D, Msyamboza KP, Oh K, Rodriguez-Artalejo F, Rojas-Martinez R, Shaw JE, Stevens GA, Tolstrup J, Zhou B, Salomon JA, Ezzati M and Danaei G. A novel risk score to predict cardiovascular disease risk in national populations (Globorisk): a pooled analysis of prospective cohorts and health examination surveys. *The lancet Diabetes & endocrinology*. 2015;3:339-55.
64. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, De Bacquer D, Ducimetiere P, Jousilahti P, Keil U, Njolstad I, Oganov RG, Thomsen T, Tunstall-Pedoe H, Tverdal A, Wedel H, Whincup P, Wilhelmsen L and Graham IM. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J*. 2003;24:987-1003.
65. van Dis I, Kromhout D, Geleijnse JM, Boer JM and Verschuren WM. Evaluation of cardiovascular risk predicted by different SCORE equations: the Netherlands as an example. *Eur J Cardiovasc Prev Rehabil*. 2010;17:244-9.
66. De Bacquer D and De Backer G. Predictive ability of the SCORE Belgium risk chart for cardiovascular mortality. *Int J Cardiol*. 2010;143:385-90.
67. Sans S, Fitzgerald AP, Royo D, Conroy R and Graham I. [Calibrating the SCORE cardiovascular risk chart for use in Spain]. *Rev Esp Cardiol*. 2007;60:476-85.
68. Usher-Smith JA, Silarova B, Schuit E, Gm Moons K and Griffin SJ. Impact of provision of cardiovascular disease risk estimates to healthcare professionals and patients: a systematic review. *BMJ open*. 2015;5:e008717.
69. Lackland DT, Elkind MS, D'Agostino R, Sr., Dhamoon MS, Goff DC, Jr., Higashida RT, McClure LA, Mitchell PH, Sacco RL, Sila CA, Smith SC, Jr., Tanne D, Tirschwell DL, Touze E and Wechsler LR. Inclusion of stroke in cardiovascular risk prediction instruments: a statement for

healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2012;43:1998-2027.

70. Hippisley-Cox J, Coupland C and Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357:j2099.
71. Harrell F. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*: Springer; 2015.
72. Steyerberg E. *Clinical prediction models: a practical approach to development, validation, and updating*: Springer Science & Business Media; 2008.
73. Steyerberg EW and Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35:1925-31.
74. Su TL, Jaki T, Hickey GL, Buchan I and Sperrin M. A review of statistical updating methods for clinical prediction models. *Stat Methods Med Res*. 2016.

Supplemental material

Contents

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 MI (recognized or unrecognized), stroke, congestive heart failure, percutaneous coronary intervention, coronary bypass surgery, or atrial fibrillation.
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, gender, 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

2.3.1 Casemix variables

For the casemix 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.

2.3.2 C-statistic

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]

2.3.3 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:

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

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:

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

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

3.3 Characteristics of included validations

Table S1: characteristics of included external validations

Reference	Validated model	Recruit years	Median FU time / Prediction horizon	Cohort	Country	Predicted outcome	N events / n particip.	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)
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 007[32]	Wilson men Total cholesterol	1995-1998	5/5	Patients from 67 health centres 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 centres 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)*
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)*
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 women	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)
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)
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 St udy	Framingham	Fatal or nonfatal CVD	539/5041	NR (40-79)	0.684 (0.012)	0.727 (0.028)
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 St udy	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 St udy	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)*
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 St udy	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 white	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)
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 women	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)
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
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	ChiNR	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	ChiNR	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 St udy	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 St udy	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
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)
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)
*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)
Simons 2003[51]	ATP III women	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 centre 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)
Yang 2016[23]	PCE men white	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)
Yang 2016[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)

Yang 2016[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)*
Yang 2016[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)
Yang 2016[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)
Yang 2016[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)*
Yang 2016[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)
Yang 2016[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)
Yang 2016[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)*
Yang 2016[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)
Yang 2016[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)
Yang 2016[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%)	6 (100%)	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%)
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 - three domains***	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	-
*low risk of bias - three domains***	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 reported because of the low number of validations. OE: observed: expected.
**Participant selection, predictors and outcome

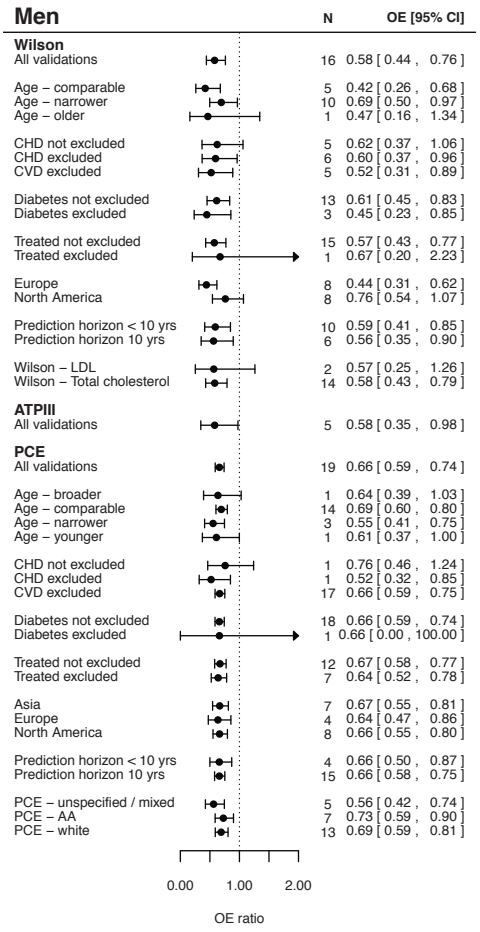
3.6 Meta-regression analyses

3.6.1 OE ratio

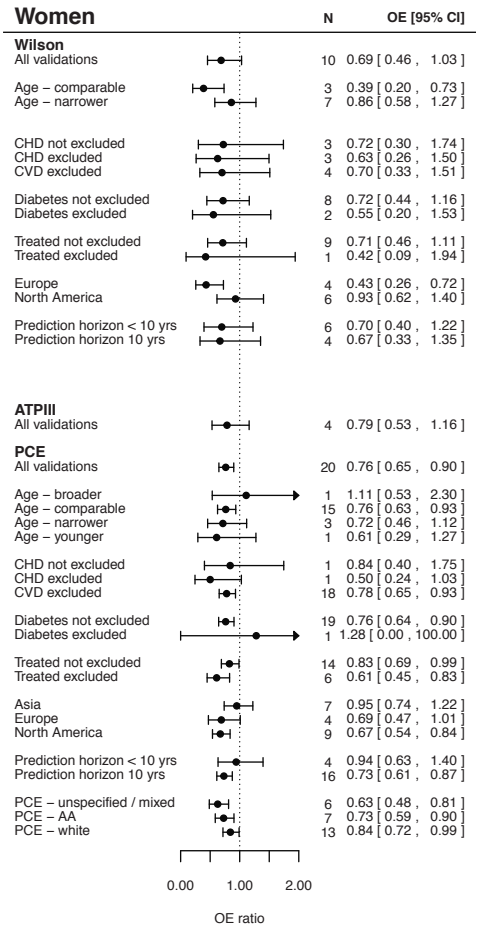
Figure S1: Results of meta-regression for the OE ratio for categorical variables (A and B) and continuous variables (C).

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.

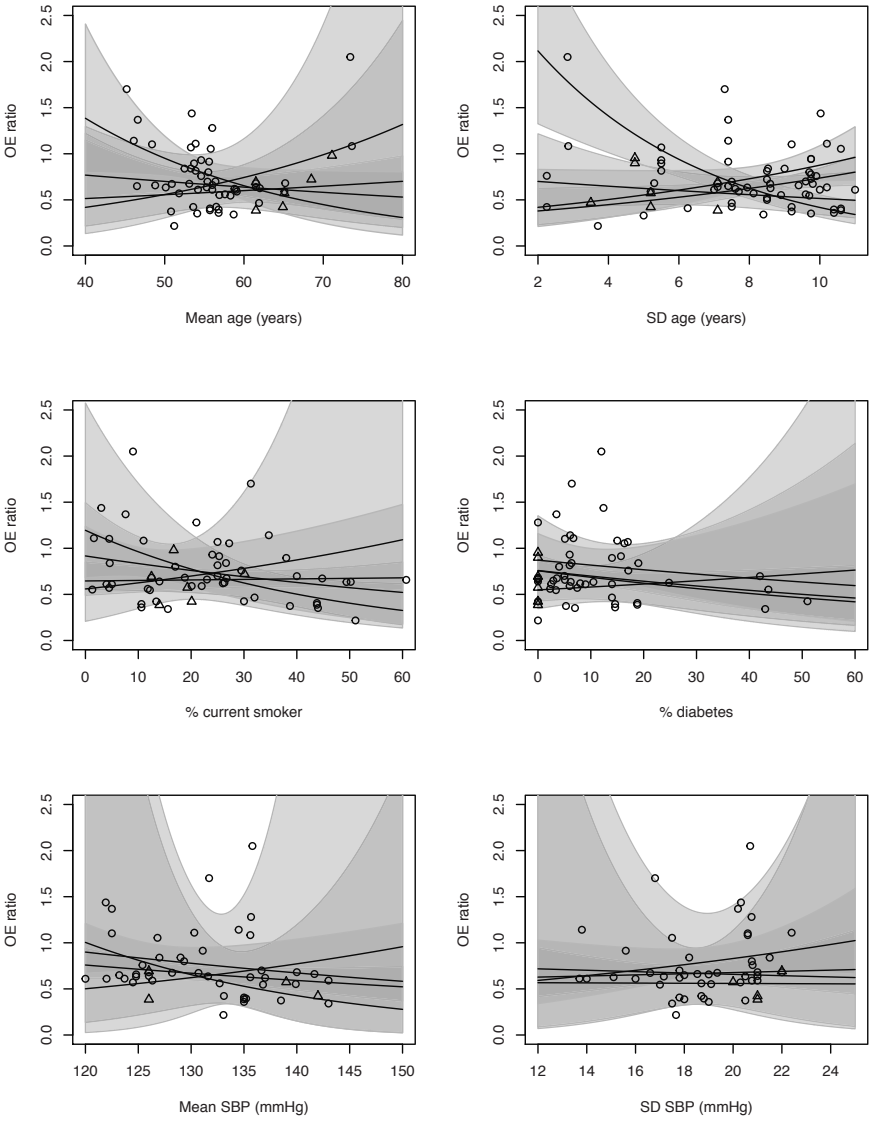
A

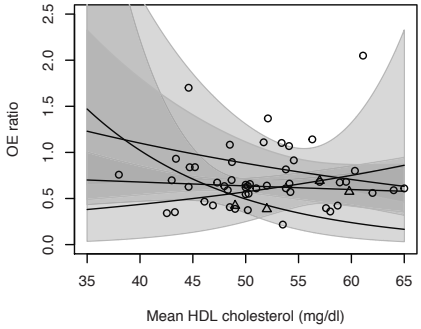


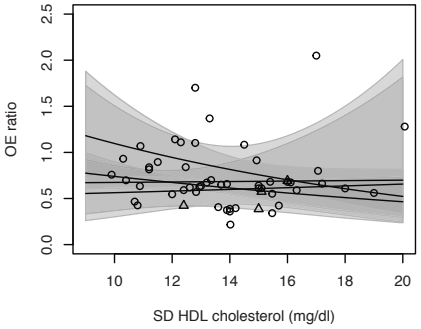
B

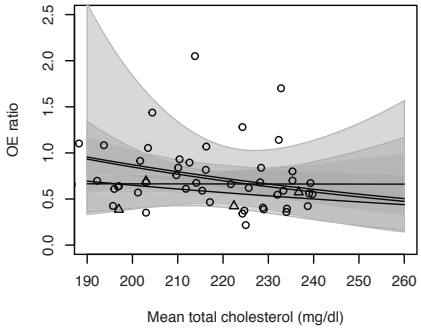


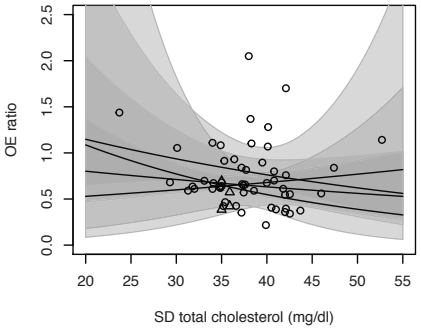
C

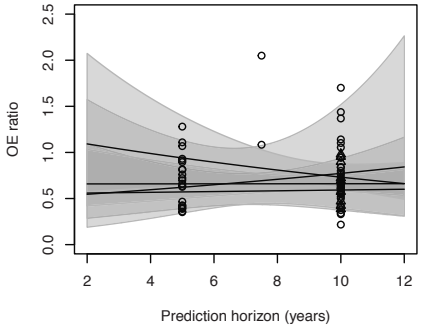

OE ratio
Mean age (years)
SD age (years)
% current smoker
% diabetes
Mean SBP (mmHg)
SD SBP (mmHg)

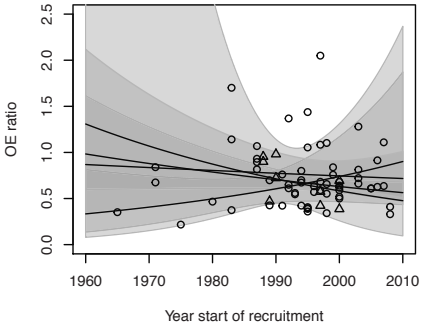

OE ratio
0.0
0.5
1.0
1.5
2.0
2.5
35
40
45
50
55
60
65
Mean HDL cholesterol (mg/dl)


OE ratio
0.0
0.5
1.0
1.5
2.0
2.5
10
12
14
16
18
20
SD HDL cholesterol (mg/dl)


OE ratio
0.0
0.5
1.0
1.5
2.0
2.5
190
200
210
220
230
240
250
260
Mean total cholesterol (mg/dl)


OE ratio
0.0
0.5
1.0
1.5
2.0
2.5
20
25
30
35
40
45
50
55
SD total cholesterol (mg/dl)


OE ratio
0.0
0.5
1.0
1.5
2.0
2.5
2
4
6
8
10
12
Prediction horizon (years)

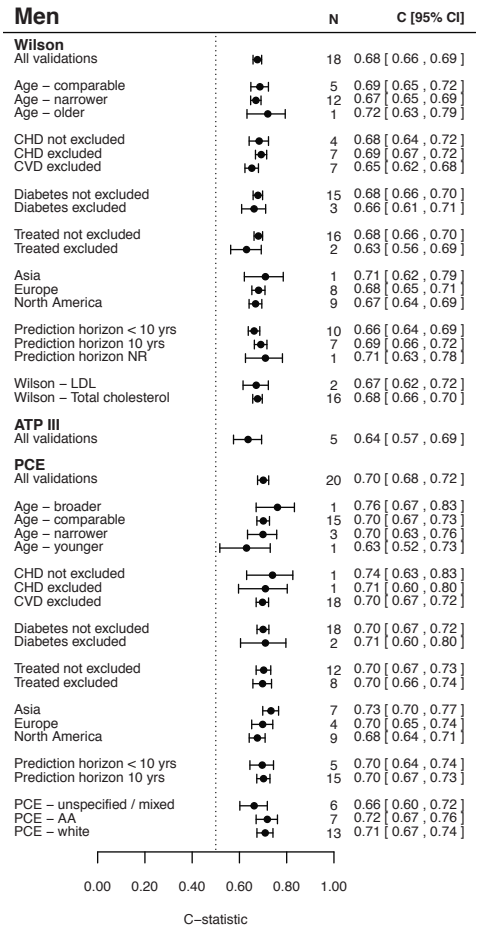

OE ratio
0.0
0.5
1.0
1.5
2.0
2.5
1960
1970
1980
1990
2000
2010
Year start of recruitment

3.6.2 C-statistic

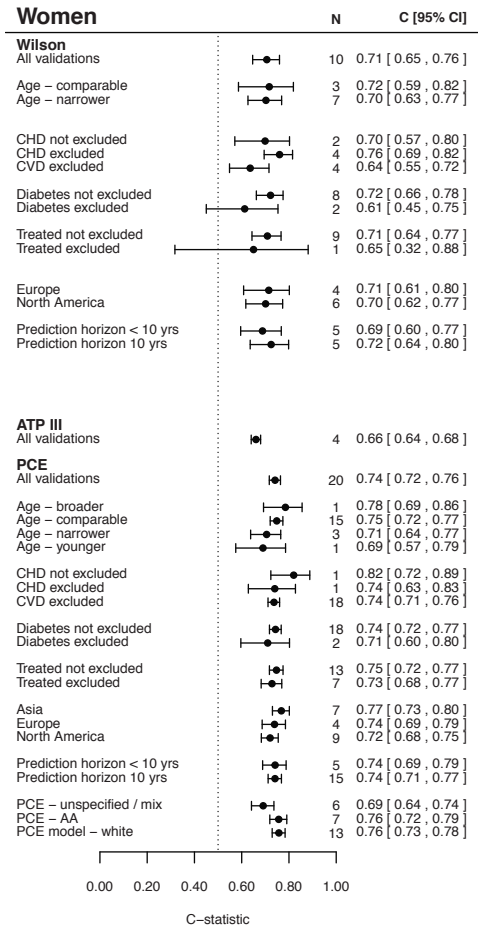
Figure S2: Results of meta-regression for the c-statistic for categorical variables (A and B) and continuous variables (C).

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.

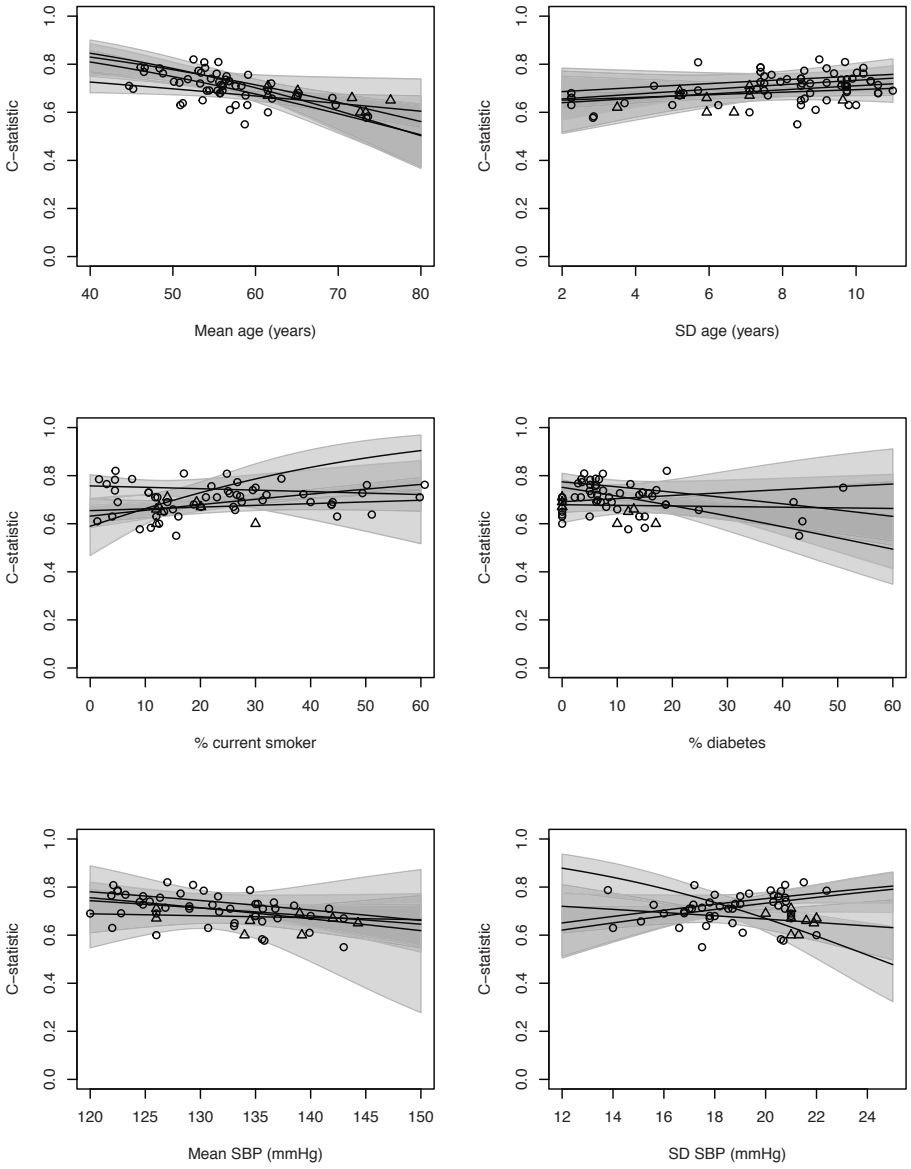
A

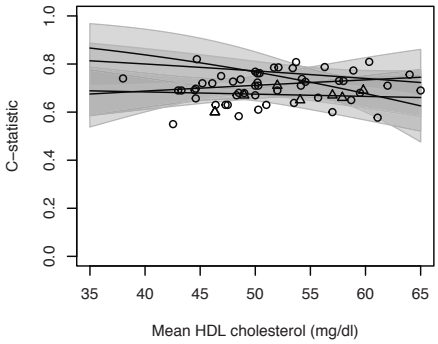


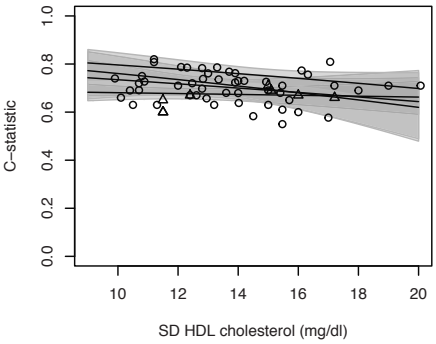
B

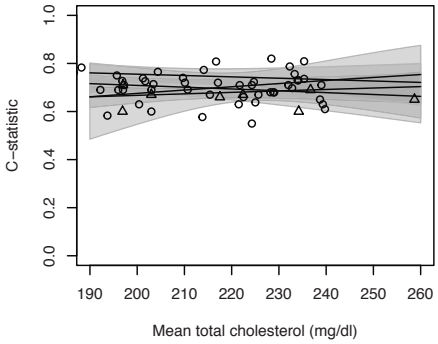


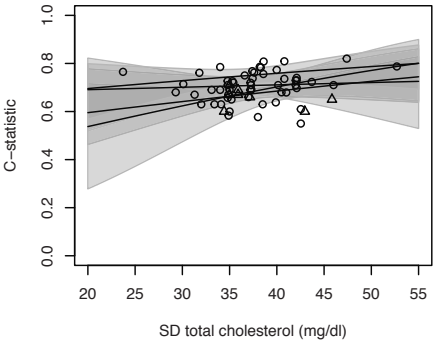
C

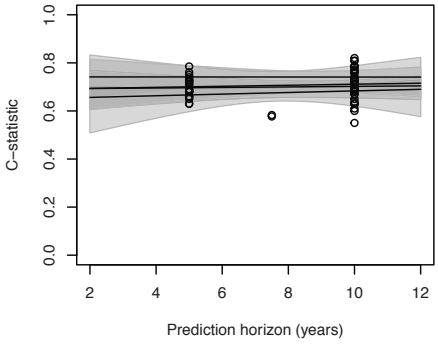

C–statistic
0.0
0.2
0.4
0.6
0.8
1.0
40
50
60
70
80
Mean age (years)
2
4
6
8
10
SD age (years)
0
10
20
30
40
50
60
% current smoker
% diabetes
120
125
130
135
140
145
150
Mean SBP (mmHg)
12
14
16
18
20
22
24
SD SBP (mmHg)


C−statistic
0.0
0.2
0.4
0.6
0.8
1.0
35
40
45
50
55
60
65
Mean HDL cholesterol (mg/dl)


C−statistic
0.0
0.2
0.4
0.6
0.8
1.0
10
12
14
16
18
20
SD HDL cholesterol (mg/dl)


C−statistic
0.0
0.2
0.4
0.6
0.8
1.0
190
200
210
220
230
240
250
260
Mean total cholesterol (mg/dl)


C−statistic
0.0
0.2
0.4
0.6
0.8
1.0
20
25
30
35
40
45
50
55
SD total cholesterol (mg/dl)


C−statistic
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
12
Prediction horizon (years)


C−statistic
0.0
0.2
0.4
0.6
0.8
1.0
1960
1970
1980
1990
2000
2010
Year start of recruitment

References

1. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H and Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97:1837-47.
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;106:3143-421.
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;285:2486-97.
4. Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC, Jr., Sorlie P, Stone NJ, Wilson PW, Jordan HS, Nevo L, Wnek J, Anderson JL, Halperin JL, Albert NM, Bozkurt B, Brindis RG, Curtis LH, DeMets D, Hochman JS, Kovacs RJ, Ohman EM, Pressler SJ, Sellke FW, Shen WK and Tomaselli GF. 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:S49-73.
5. Wan X, Wang W, Liu J and 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 and Tilley BC. A coronary heart disease risk score based on patient-reported information. *Am J Cardiol*. 2007;99:1236-41.
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;25:559-73.
8. Hanley JA and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29-36.
9. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, Riley RD and Moons KG. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356:i6460.
10. IntHout J, Ioannidis JP and Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol*. 2014;14:25.
11. Pennells L, Kaptoge S, White IR, Thompson SG and Wood AM. Assessing risk prediction models using individual participant data from multiple studies. *Am J Epidemiol*. 2014;179:621-32.
12. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG and Riley RD. 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 and Tai ES. The metabolic syndrome and mortality: the Singapore Cardiovascular Cohort Study. *Clin Endocrinol (Oxf)*. 2008;69:225-30.
14. Stork S, Feelders RA, van den Beld AW, Steyerberg EW, Savelkoul HFJ, Lamberts SWJ, Grobbee DE and Bots ML. Prediction of mortality risk in the elderly. *Am J Med*. 2006;119:519-25.

15. Barroso LC, Muro EC, Herrera ND, Ochoa GF, Hueros JIC and 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;28:242-8.
16. Ridker PM, Buring JE, Rifai N and Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. *JAMA*. 2007;297:611-9.
17. Berry JD, Lloyd-Jones DM, Garside DB and Greenland P. Framingham risk score and prediction of coronary heart disease death in young men. *Am Heart J*. 2007;154:80-6.
18. Dunder K, Lind L, Zethelius B, Berglund L and 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;148:596-601.
19. Jung KJ, Jang Y, Oh DJ, Oh BH, Lee SH, Park SW, Seung KB, Kim HK, Yun YD, Choi SH, Sung J, Lee TY, Kim SH, Koh SB, Kim MC, Chang Kim H, Kimm H, Nam C, Park S and Jee SH. The ACC/AHA 2013 pooled cohort equations compared to a Korean Risk Prediction Model for atherosclerotic cardiovascular disease. *Atherosclerosis*. 2015;242:367-375.
20. DeFilippis AP, Young R, Carrubba CJ, McEvoy JW, Budoff MJ, Blumenthal RS, Kronmal RA, McClelland RL, Nasir K and Blaha MJ. An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort. *Ann Intern Med*. 2015;162:266-75.
21. De Filippis AP, Young R, McEvoy JW, Michos ED, Sandfort V, Kronmal RA, McClelland RL and Blaha MJ. 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;38:598-608.
22. Muntner P, Colantonio LD, Cushman M, Goff DC, Jr., Howard G, Howard VJ, Kissela B, Levitan EB, Lloyd-Jones DM and Safford MM. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *JAMA*. 2014;311:1406-15.
23. Yang X, Li J, Hu D, Chen J, Li Y, Huang J, Liu X, Liu F, Cao J, Shen C, Yu L, Lu F, Wu X, Zhao L, Wu X and Gu D. Predicting the 10-Year Risks of Atherosclerotic Cardiovascular Disease in Chinese Population: The China-PAR Project (Prediction for ASCVD Risk in China). *Circulation*. 2016;134:1430-1440.
24. Mortensen MB, Afzal S, Nordestgaard BG and Falk E. Primary Prevention With Statins: ACC/AHA Risk-Based Approach Versus Trial-Based Approaches to Guide Statin Therapy. *J Am Coll Cardiol*. 2015;66:2699-2709.
25. Mortensen MB, Nordestgaard BG, Afzal S and 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;38:586-594.
26. D'Agostino RB, Sr., Grundy S, Sullivan LM and Wilson P. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA*. 2001;286:180-7.

27. Koller MT, Leening MJG, Wolbers M, Steyerberg EW, Hunink MGM, Schoop R, Hofman A, Bucher HC, Psaty BM, Lloyd-Jones DM and Witteman JCM. 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;157:389-97.
28. Koller MT, Steyerberg EW, Wolbers M, Stijnen T, Bucher HC, Hunink MGM and Witteman JCM. Validity of the Framingham point scores in the elderly: results from the Rotterdam study. *Am Heart J*. 2007;154:87-93.
29. Andersson C, Enserro D, Larson MG, Xanthakis V and 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. *J Am Heart Assoc*. 2015;4.
30. Buitrago F, Calvo-Hueros JI, Canon-Barroso L, Pozuelos-Estrada G, Molina-Martinez L, Espigares-Arroyo M, Galan-Gonzalez JA and Lillo-Bravo FJ. Original and REGICOR Framingham functions in a nondiabetic population of a Spanish health care center: a validation study. *Ann Fam Med*. 2011;9:431-8.
31. Chia YC, Lim HM and 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, Cordon F, Grau M, Cabre-Vila JJ and Marrugat J. Estimating cardiovascular risk in Spain using different algorithms. *Rev Esp Cardiol*. 2007;60:693-702.
33. Cook NR and 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;174:1964-71.
34. Cooper JA, Miller GJ and 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;181:93-100.
35. De Las Heras Gala T, Geisel MH, Peters A, Thorand B, Baumert J, Lehmann N, Jockel KH, Moebus S, Erbel R, Mahabadi AA, Koenig W, Heinrich J, Holle R, Leidl R, Meisinger C, Strauch K, Roggenbuck U, Slomiany U, Beck EM, Offner A, Munkel S, Bauer M, Schrader S, Peter R and Hirche H. Recalibration of the ACC/AHA risk score in two population-based German cohorts. *PLoS One*. 2016;11 (10) (no pagination):e0164688.
36. Emdin CA, Khera AV, Natarajan P, Klarin D, Baber U, Mehran R, Rader DJ, Fuster V and Kathiresan S. Evaluation of the Pooled Cohort Equations for Prediction of Cardiovascular Risk in a Contemporary Prospective Cohort. *Am J Cardiol*. 2017;119:881-885.
37. Empana JP, Ducimetiere P, Arveiler D, Ferrieres J, Evans A, Ruidavets JB, Haas B, Yarnell J, Bingham A, Amouyel P, Dallongeville J and Group PS. Are the Framingham and PROCAM coronary heart disease risk functions applicable to different European populations? The PRIME Study. *Eur Heart J*. 2003;24:1903-11.
38. Ferrario M, Chiodini P, Chambless LE, Cesana G, Vanuzzo D, Panico S, Sega R, Pilotto L, Palmieri L, Giampaoli S and Group CPR. Prediction of coronary events in a low incidence population. Assessing accuracy of the CUORE Cohort Study prediction equation. *Int J Epidemiol*. 2005;34:413-21.

39. Jee SH, Jang Y, Oh DJ, Oh BH, Lee SH, Park SW, Seung KB, Mok Y, Jung KJ, Kimm H, Yun YD, Baek SJ, Lee DC, Choi SH, Kim MJ, Sung J, Cho B, Kim ES, Yu BY, Lee TY, Kim JS, Lee YJ, Oh JK, Kim SH, Park JK, Koh SB, Park SB, Lee SY, Yoo CI, Kim MC, Kim HK, Park JS, Kim HC, Lee GJ and Woodward M. A coronary heart disease prediction model: The Korean heart study. *BMJ Open*. 2014;4.
40. Kavousi M, Leening MJ, Nanchen D, Greenland P, Graham IM, Steyerberg EW, Ikram MA, Stricker BH, Hofman A and Franco OH. 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;311:1416-23.
41. Khalili D, Asgari S, Hadaegh F, Steyerberg EW, Rahimi K, Fahimfar N and Azizi F. 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;184:587-594.
42. Lee CH, Woo YC, Lam JKY, Fong CHY, Cheung BMY, Lam KSL and Tan KCB. Validation of the Pooled Cohort equations in a long-term cohort study of Hong Kong Chinese. *J Clin Lipidol*. 2015;9:640-646.
43. Lloyd-Jones DM, Wilson PWF, Larson MG, Beiser A, Leip EP, D'Agostino RB and Levy D. Framingham risk score and prediction of lifetime risk for coronary heart disease. *Am J Cardiol*. 2004;94:20-4.
44. Marrugat J, Subirana I, Comin E, Cabezas C, Vila J, Elosua R, Nam BH, Ramos R, Sala J, Solanas P, Cordon F, Gene-Badia J and D'Agostino RB. Validity of an adaptation of the Framingham cardiovascular risk function: the VERIFICA Study. *J Epidemiol Community Health*. 2007;61:40-7.
45. Pike MM, Decker PA, Larson NB, St Sauver JL, Takahashi PY, Roger VL, Rocca WA, Miller VM, Olson JE, Pathak J and Bielinski SJ. Improvement in Cardiovascular Risk Prediction with Electronic Health Records. *J Cardiovasc Transl Res*. 2016;9:214-222.
46. Rana JS, Tabada GH, Solomon MD, Lo JC, Jaffe MG, Sung SH, Ballantyne CM and Go AS. Accuracy of the Atherosclerotic Cardiovascular Risk Equation in a Large Contemporary, Multiethnic Population. *J Am Coll Cardiol*. 2016;67:2118-2130.
47. Reissigova J and Zvarova J. The Framingham risk function underestimated absolute coronary heart disease risk in Czech men. *Methods Inf Med*. 2007;46:43-9.
48. Rodondi N, Locatelli I, Aujesky D, Butler J, Vittinghoff E, Simonsick E, Satterfield S, Newman AB, Wilson PWF, Pletcher MJ, Bauer DC and Health ABCS. Framingham risk score and alternatives for prediction of coronary heart disease in older adults. *PLoS One*. 2012;7:e34287.
49. Ryckman EM, Summers RM, Liu J, Munoz del Rio A and Pickhardt PJ. Visceral fat quantification in asymptomatic adults using abdominal CT: is it predictive of future cardiac events? *Abdom Imaging*. 2015;40:222-6.
50. Simmons RK, Sharp S, Boekholdt SM, Sargeant LA, Khaw K-T, Wareham NJ and Griffin SJ. 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;168:1209-16.

51. Simons LA, Simons J, Friedlander Y, McCallum J and Palaniappan L. Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study. *Med J Aust*. 2003;178:113-6.
52. Suka M, Sugimori H and Yoshida K. Application of the updated Framingham risk score to Japanese men. *Hypertens Res*. 2001;24:685-9.
53. Sussman JB, Wiitala WL, Zawistowski M, Hofer TP, Bentley D and Hayward RA. The Veterans Affairs Cardiac Risk Score: Recalibrating the Atherosclerotic Cardiovascular Disease Score for Applied Use. *Med Care*. 2017;55:864-870.
54. Vaidya D, Yanek LR, Moy TF, Pearson TA, Becker LC and 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;100:1410-5.
55. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW and Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2014.

CHAPTER 3

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

Published
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 Dec;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), the timing of their use, and the handling of such treatment use in the analyses were extracted and summarised.

Results

302 articles were included in the review. Treatment use was not mentioned in 91 (30%) articles. One hundred forty-six (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. Seventy-nine 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.

Introduction

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 handling of treatment use in cardiovascular prognostic model research. We conclude with suggestions for the handling and reporting of treatment use in prognostic model research.

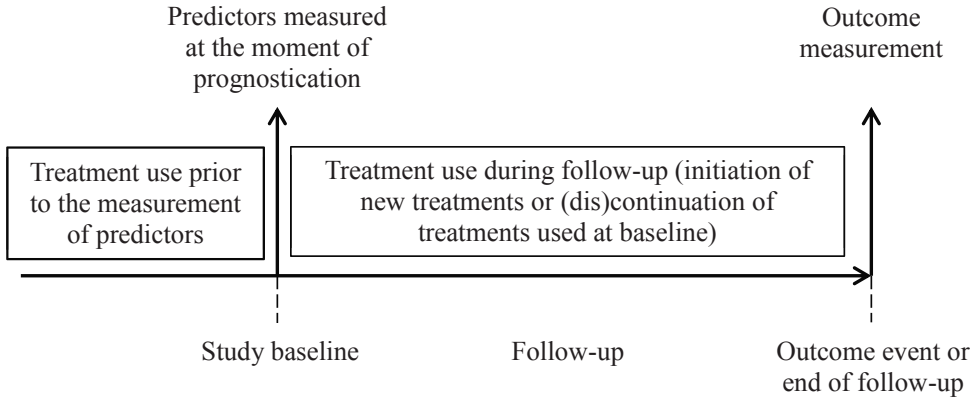
Methods

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 summarise, 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 this 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[8,11] and further guidance can be found in Table 1 (see below).

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]

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
Design of study used for prognostic modelling	
Observational	286
Randomised trial	16
Follow-up period (years)	10, (6, 12); 15%[a]
Prediction horizon (years)	10, (8, 10); 12%[a]

* 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

[a] Values represent as follows: median (lower quartile, upper quartile), percentage of studies that did not report this information

Data sources, search, and study selection

In brief, a search was performed on 1 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 (Supplement 1) 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 Supplement 1).

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 (Supplement 2). A summary of the article inclusion process is presented in Figure 2. 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 2.

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 Tables 3 and 4, respectively.

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%). Twenty articles (15%) excluded treated individuals from the analysis. Changes in treatment use during follow-up were not modelled.

Figure 2: A flow diagram of article inclusion and exclusion

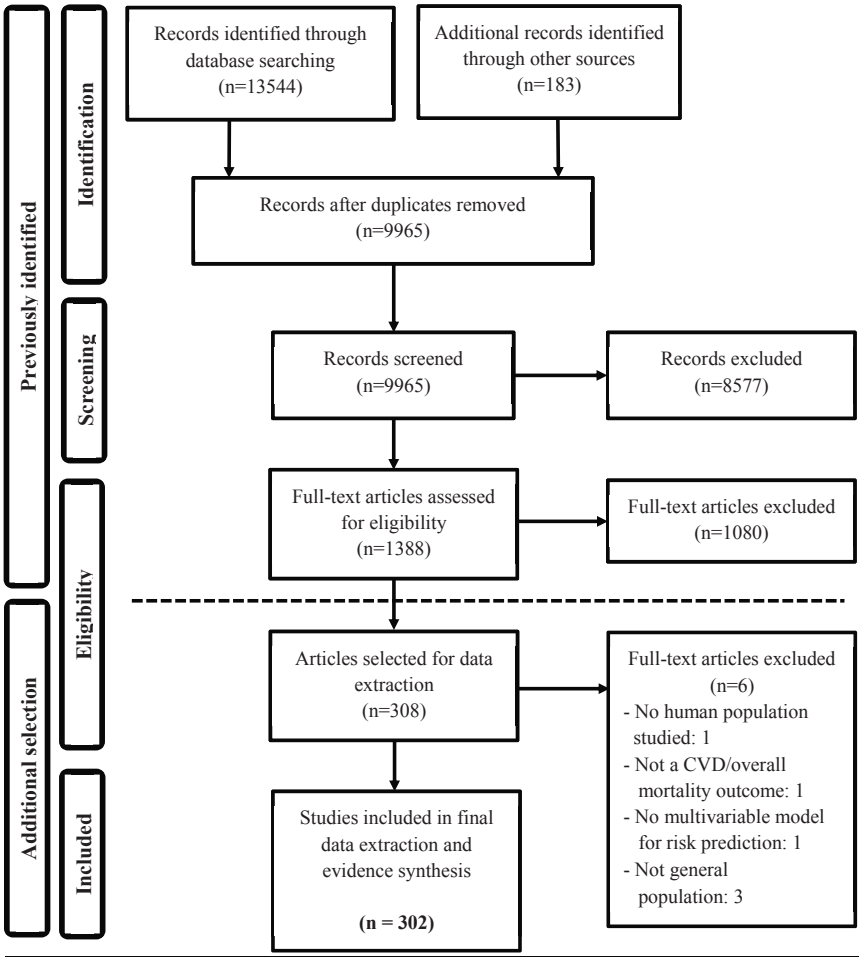
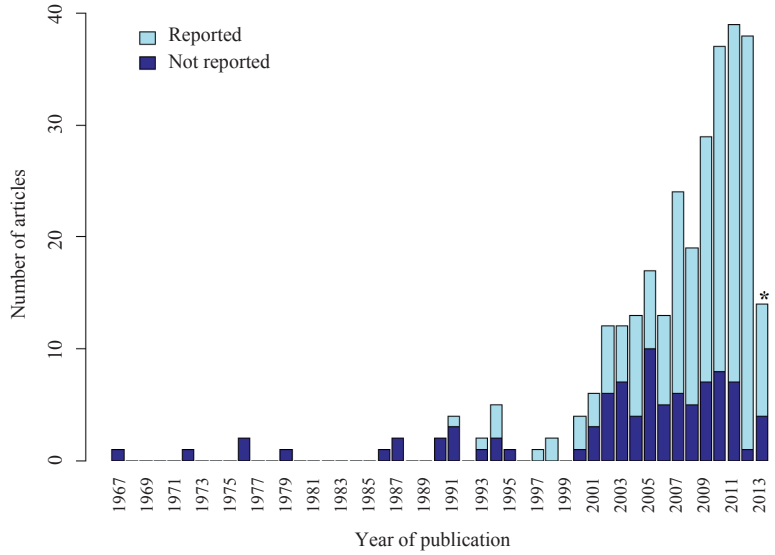


Figure 3: Reporting of treatment in CVD prognostic modelling studies over time.

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.

Table 2. Reporting of treatment use by study type.

Reported treatment	Overall (*n*=302) (%)[a]	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)

[a] 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	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)
Exclusion as a sensitivity analysis	9 (7)	5 (4)	4 (3)
Stratification by treatment	1 (1)	0	0
Treatment included in the outcome	23 (19)	58 (43)	35 (24)
Treatment modelled as a predictor	54 (44)	48 (59)[a]	–
Antihypertensives	49 (40)	44 (54)[a]	–
Lipid-lowering agents	12 (10)	15 (11)[a]	–
Other medications	2 (2)	5 (4)[a]	–
Lifestyle interventions	1 (1)	0[a]	–
Surgical interventions	0	0[a]	–
Modelled directly (not a composite[b])	37 (30)	44 (54)[a]	–
Baseline treatment modelled	41 (33)	36 (44)[a]	–
Changes in treatment during follow-up modelled	0	0[a]	–
Treatment at end of study modelled	0	1 (1)[a]	–
Not clearly reported	12 (10)	11 (8)	–
Treatment interactions considered	21 (17)	7 (5)[a]	–

[a]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 bComposite predictors are here defined as the combination of two or more variables (including treatment use) into a single predictor

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%). Fifty-three 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.

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 4 years include the ACC/AHA Pooled Cohort equation,[16] the Globorisk CVD assessment tool,[19] and the Qrisk-3 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 (Table 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: "What do we mean by "treatment" and when is it a problem?"). 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 on 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, Table 1 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.

Table 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.

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. from 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, 25, 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.

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 recognised 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.

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. National Institute for Health and Care Excellence. Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. NICE Clinical Guideline 181. 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 SM, Glasziou P. Risk prediction continue to ignore treatment effects. BMJ: British Medical Journal Rapid Responses; 2010.
8. Groenwold RH, Moons KG, Pajouheshnia R, Altman DG, Collins GS, Debray TP, et al. Explicit inclusion of treatment in prognostic modelling was recommended in observational and randomised settings. J Clin Epidemiol. 2016.
9. Peek N, Sperrin M, Mamas M, Van Staa T, Buchan I. Hari Seldon, QRISK3, and the Prediction Paradox. BMJ; 2017.
10. Grobbee DE, Hoes AW. Clinical Epidemiology. 2nd ed: Jones & Bartlett Publishers; 2014.
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 medicine. 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.
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. BMJ. 2015;350:g7594.
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. Annals of internal medicine. 2015;162(1):W1-W73.
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. 2015;68(3):279-89.
25. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, 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.
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

Supplement 1. List of items for data extraction

1. General study information

i) General study aims.

ii) Study type.

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

iv) Study design.

v) Start of data collection.

vi) End of data collection.

vii) Length of follow-up.

viii) Intended prediction horizon.

2. Reporting of treatment-specific information

i) Where in the article is information about treatment reported?

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

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

iv) 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.

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

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

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

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

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

x) 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? Give details.

3. Accounting for treatment use in the analysis

i) If treatments are not accounted for in the analysis, is a reason given for why this is so?

- If so, give details.

ii) 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?

iii) 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)? Give details.
- Are treatment interactions with other variables modelled?
- Is treatment included as a predictor in the final model?, If not, why not?
- Is a treatment modelled alongside any associated condition (i.e. blood pressure-lowering medication and blood pressure)?

iv) Are analyses stratified according to treatment use?

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

Supplement 2. 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 4

Accounting for treatment use when validating a prognostic model: a simulation study

Published
Pajouheshnia R, Peelen LM, Moons KG, Reitsma JB, Groenwold RH. Accounting for treatment use when validating a prognostic model: a simulation study. BMC medical research methodology. 2017 Dec;17(1):103.

Abstract

Background

Prognostic models often show poor performance when applied to independent validation data sets. We illustrate how treatment use in a validation set can affect measures of model performance and present the uses and limitations of available analytical methods to account for this using simulated data.

Methods

We outline how the use of risk-lowering treatments in a validation set can lead to an apparent overestimation of risk by a prognostic model that was developed in a treatment-naïve cohort to make predictions of risk without treatment. Potential methods to correct for the effects of treatment use when testing or validating a prognostic model are discussed from a theoretical perspective. Subsequently, we assess in simulated data sets, the impact of excluding treated individuals and the use of inverse probability weighting (IPW) on the estimated model discrimination (c-index) and calibration (observed:expected ratio and calibration plots) in scenarios with different patterns and effects of treatment use.

Results

Ignoring the use of effective treatments in a validation data set leads to poorer model discrimination and calibration than would be observed in the untreated target population for the model. Excluding treated individuals provided correct estimates of model performance only when treatment was randomly allocated, although this reduced the precision of the estimates. When the assumptions of IPW were met, IPW followed by exclusion of the treated individuals provided correct estimates of model performance in data sets where treatment use was either random or moderately associated with an individual's risk. IPW followed by exclusion yielded incorrect estimates in the presence of non-positivity or an unobserved confounder.

Conclusions

When validating a prognostic model developed to make predictions of risk without treatment, treatment use in the validation set can bias estimates of the performance of the model in future targeted individuals, and should not be ignored. When treatment use is random, treated individuals can be excluded from the analysis. When treatment use is non-random, IPW followed by the exclusion of treated individuals is recommended, however, this method is sensitive to violations of its assumptions.

Introduction

Prognostic models have a range of applications, from risk stratification, to use in making individualized predictions to help counsel patients or guide healthcare providers when deciding whether or not to recommend a certain treatment or intervention.[1-3] Before prognostic models can be used in practice, their predictive performance (e.g. discrimination and calibration)- in short, performance - should be evaluated in a set of individuals who are representative of future targeted individuals. In studies that use independent data to validate a previously developed prognostic model, performance is often considerably worse than in the development set.[4] This may be due to, for example, overfitting of the model in the development data set[5, 6] or differences in case-mix (between the development set and validation sets.[7-10]

One aspect that can vary considerably between data sets used for model development and validation is the use of treatments or preventative interventions that affect (reduce) the occurrence of the outcomes under prediction. Although a difference in the use of treatments between a development and validation set is generally viewed as a difference in case-mix characteristics, treatment use in a validation set can actually lead to further problems. When additional treatment use in a validation set (compared to the development set) results in a markedly lower incidence of the outcome under prediction, the predictive performance of the model will likely be affected. A challenge arises when a prognostic model has originally been developed in order to make predictions of "untreated risks", i.e. predictions of an individual's prognosis without certain treatments, to guide the decision to initiate those treatments in future targeted individuals. Ideally these models should be validated in data sets in which individuals remain untreated with those specific treatments throughout follow-up - so-called treatment-naïve populations. However, the use of such treatment-naïve populations is uncommon and poor performance of a prognostic model seen in a validation study could be directly attributed to treatment use in the validation data set.[11, 12]

Ignoring the effects of treatment use in the development phase of a prognostic model for the prediction of untreated risks has already been shown to lead to a model that underestimates this risk in future targeted individuals.[13] However, it is not clear to what extent treatment use in a validation set might influence the observed performance of a prognostic model that was developed in a treatment-naïve population. In addition, how to account for treatment use in a treated validation set in order to correctly estimate how a prognostic model would perform in its target (untreated) population, remains unclear.

In this paper, we provide a detailed explanation of when and how treatment use in a validation set can bias the estimation of the performance of a prognostic model in

future targeted (untreated) individuals and compare different analytical approaches to correctly estimate the performance of a model using a partly treated validation data set in a simulation study.

Methods

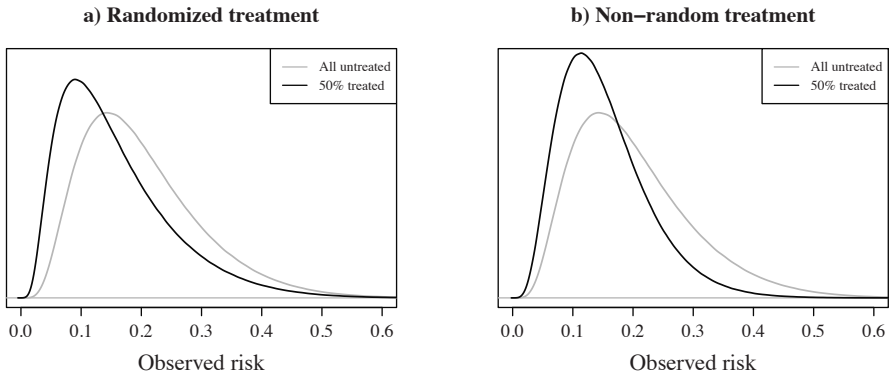
Problems with ignoring treatment use in a validation study

If individuals in a validation set receive an effective treatment during follow-up, their risk of developing the outcome will decrease. Figures 1a and b show the effect of treatment use on the distribution of risks in data sets that represent data from a randomized trial (RCT) and a non-randomized study (e.g. routine care data or data from an observational cohort study) in which treatment use was more likely in high-risk individuals. In the event of the use of an effective treatment, fewer individuals will develop the outcome than would have, had they remained untreated, and thus the observed outcome frequencies will be lower than the predicted "untreated" outcome frequencies. As a result, a prognostic model developed for making predictions of risk without that treatment (i.e. models used to guide the initiation of a certain treatment) will erroneously appear to overestimate risk in a partially treated validation set, regardless of how treatments have been allocated. As the aim, in this case, is to estimate the performance of the model when used for future, untreated individuals, measures of model discrimination and calibration will give a biased representation of the performance of the model when used in practice for making untreated outcome predictions, if treatment use in the validation set is ignored.

The effect that treatment use will have on measures of model performance in a validation study will depend on a number of factors, including the strength of the effect of treatment on the outcome risk, the proportion of individuals receiving treatment, and the underlying pattern of treatment use. If a treatment has a weak effect on the outcome risk or only a small proportion of individuals are treated in a validation set, the impact on model discrimination and calibration will be relatively small. Furthermore, the way in which treatments are allocated to individuals, whether treatment is allocated randomly, as in data from a RCT, or non-randomly and treatment use is rather based on an individual's risk-profile or according to strict treatment guidelines, will influence the impact that treatment use will have in a validation study. If, for example, high-risk individuals are selectively treated, we can anticipate an even greater impact of treatment use on measures of model performance. In this case, the distribution of observed risks will become narrower, due to the risk-lowering effects of treatment in the high-risk individuals (see Figure 1b), making it more difficult for the model to discriminate between individuals who will or will not develop the outcome, and the calibration in high-risk individuals will be most greatly affected.

Figure 1 a-b: Risk distributions in two simulated validation sets.

50% of individuals received an effective treatment (relative odds reduction on treatment: 0.5), (see Table 2 scenarios 2 and 1, respectively, for details). a: the model was validated on the combined treatment and control group of a randomised trial. b: the model was validated using data from a non-randomised setting where the probability of receiving treatment depended on an individual's (untreated) outcome risk. Black lines represent the observed risks in the validation set, after treatment. Grey lines represent the risks of the same individuals had they (hypothetically) remained untreated



Methods to account for treatment use

In this section we describe possible approaches to account for treatment use in a validation study. For each method, the rationale, expected result of its use, and potential issues are outlined. A summary of the methods, including additional technical details can be found in Table 1.

Table 1: Possible methods to account for the effects of treatment in a validation set.

Approach	Implementation	Key considerations
1. Exclude treated individuals	1. Exclude any individual who received treatment between the point of prediction and the assessment of the outcome from the analysis. 2. Estimate model performance in only the untreated subset.	- Provides correct estimates of performance in the (untreated) target population if treatment use is not associated with other prognostic factors.† - Decreases the effective sample size.
2. Inverse probability weighting	1. Fit a propensity score (PS) model for treatment in the validation set using logistic regression: $\text{logit}(\text{Tr}_i) = \alpha_0 + \sum_{i=1}^{n}(\alpha_i X_i)$ 2. Calculate PS for individuals using the estimates from the fitted PS model: $PS_i = \sum_{i=1}^{n}(\hat{\alpha}_i X_i)$ 3. Calculate inverse probability weights (w_i) for each untreated individual based on their individual PS: $w_i = 1 / (1 - PS_i)$.[14] 4. Exclude treated individuals from the analysis set. 5. *(optional)* Truncate weights.[15] 6. Estimate weighted measures of model performance in only the untreated subset.	- Provides correct estimates of performance in (untreated) target population if treatment use is or is not associated with other prognostic factors, provided key assumptions of IPW are met.† - Does not provide correct estimates in the presence of non-positivity, or when there are unobserved predictors that are strongly associated with both the outcome and use of treatment.[16, 17] - Exclusion of treated individuals decreases the effective sample size. - Extreme weights can further reduce precision and introduce bias.
3. Recalibration	1. Calculate the linear predictor of the prognostic model: $LP0_i = \sum_{i=1}^{n}(\hat{\beta}_i X_i)$ *2.*Re-estimate the model intercept in the full validation data.[18, 19] $\text{logit}(Y_i) = \gamma_0 + \textit{offset}(LP0_i)$ 3. Calculate the updated linear predictor. $LP1_i = \hat{\gamma}_0 + LP0_i$ 4. Estimate model performance using LP1.	- Does not affect discrimination. - Not sufficient to correct calibration if relative treatment effects are heterogeneous or use is associated with an individual's risk. - Adjusts for other differences in case-mix leading to misleading estimates of the calibration of the original model.

4. Model treatment	1. Refit the original prognostic model using the full validation data, including an indicator term for treatment use and treatment interaction terms. i) with recalibration of the intercept: $\text{logit}(Y_i) = \gamma_0 + \textit{offset}(\text{LP0}_i) + \gamma_{Tr}\,\text{Tr}_i$ * ii) with a full refit of the original model: $\text{logit}(Y_i) = \gamma_0 + \sum_{i=1}^{n}(\gamma_i X_i) + \gamma_{Tr}\text{Tr}_i$ * 2. Calculate the updated linear predictor. i) $\text{LP2}_i = \hat{\gamma}_0 + \sum_{i=1}^{n}(\hat{\beta}_i X_i) + \hat{\gamma}_{Tr}\text{Tr}_i$ * ii) $\text{LP3}_i = \hat{\gamma}_0 + \sum_{i=1}^{n}(\hat{\gamma}_i X_i) + \hat{\gamma}_{Tr}\text{Tr}_i$ * 3. Estimate model performance using LP2 or LP3.	- Can lead to an over-estimation of model discrimination. - Adjusts for other differences in case-mix leading to misleading estimates of the calibration of the original model.

Abbreviations: Xi: design matrix (predictor values) for individual i; Yi: outcome for individual i; LP: linear predictor; PS: propensity score; Tr: treatment.
$\hat{\alpha}_i$ represent coefficients of the treatment propensity model for individual i.
$\hat{\beta}_i$ represent coefficients of the original prognostic model for individual i.
$\hat{\gamma}_i$ represent coefficients of the updated prognostic model for individual i.
* Interaction terms between treatment use and predictors should be included where necessary.
† Estimates will be correct providing all other modelling assumptions are met.

Exclusion of treated individuals from the analysis

A common and straightforward approach to remove the effects of treatment is to exclude from the analysis individuals in the validation data set who received treatment. In doing this, one assumes that the untreated subset will resemble the untreated target population for the model.

As Figure 2a shows, in settings where treatment is randomly allocated (Table 2, scenario 2), the exclusion of treated individuals will result in a validation set that is indeed still representative of the target population. As a result, measures of discrimination and calibration are the same as they would be had all individuals remained untreated, and thus are correct estimates of the performance of the model in its target population. However, the effective sample size is reduced, (e.g. a 50% reduction in the case of a RCT with 1:1 randomization).

Figure 2b represents a study where treatment allocation was non-random and high-risk individuals had a higher probability of being treated (Table 2, scenario 1). If treatments were initiated between the moment of making a prediction and the assessment of the outcome, the exclusion of treated individuals results in a subset of individuals with a lower risk on average than in the untreated target population. As a result, the case-mix (in terms of risk profile) in the data set will become more homogenous, and one can expect measures of discrimination to decrease,[9, 14] underestimating the true

discriminative ability of the model in future targeted individuals. While this approach may appear to provide correct estimates of calibration, the interpretation of these measures is limited due to the inherent selection bias. The non-randomly untreated individuals only represent a portion of the total target population. Hence, estimates of model performance may provide little information about how well calibrated the model is for high-risk individuals, as these have been actively excluded.

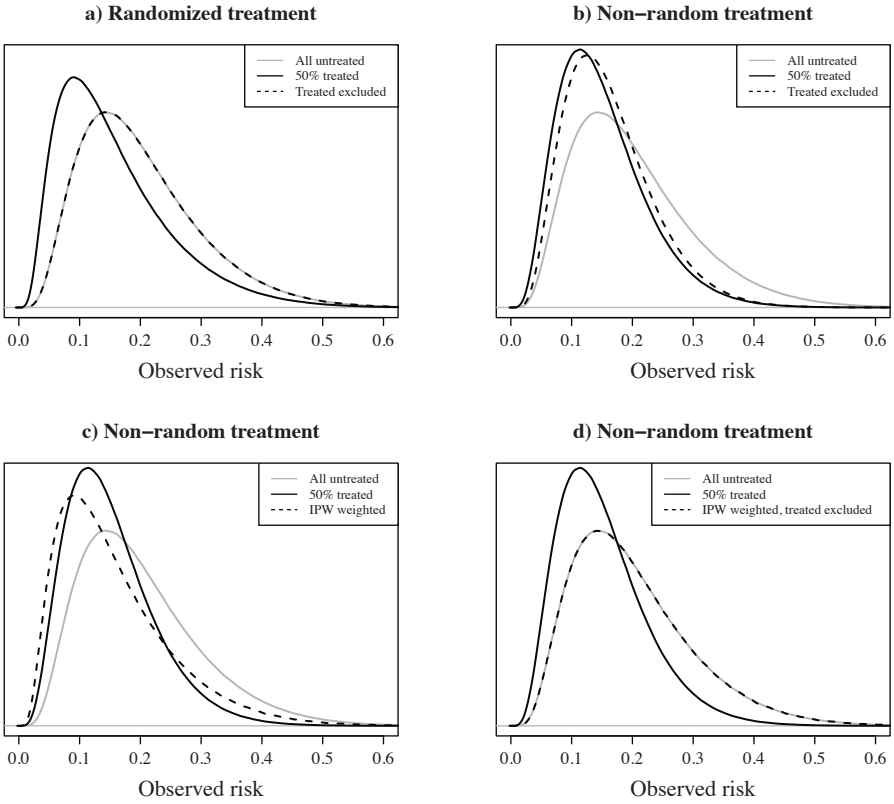
Inverse probability weighting

An alternative approach for model validation in data sets with non-random treatment use would be to balance the data in such a way that it resembles that of a RCT. Inverse probability weighting (IPW) is a method applied in studies where the aim is to obtain an estimate of the causal association between an exposure and outcome, accounting for the influence of confounding variables on the effect estimate.[15] A "treatment propensity model" is first fitted to the validation data, regressing an indicator (yes/no) of treatment use (dependent variable) on any measured variables that may be predictive of treatment use (independent variables), including the predictors of the prognostic model that is being evaluated.[16] Subsequently this treatment propensity model is then used to estimate for each individual in the validation set the probability of receiving the treatment, based on his/her observed variables (risk profile). Following this, each individual is weighted by the inverse of their own probability of the actual treatment received,[17] resulting in a distribution of risks in the validation set that resembles what would have been seen had treatments been randomly allocated, as shown by the similarity of the solid black line in Figure 2a and the dashed black line in Figure 2c. By excluding treated individuals after deriving weights, the resulting validation set should resemble the untreated target population, as seen in Figure 2d. However, this will again result in a smaller effective sample size for the validation.

IPW is subject to a number of theoretical assumptions.[15, 18, 19] One example of a violation of these assumptions is practical non-positivity (i.e. it may be that in some risk strata no subjects received the treatment),[20] which may arise if a subset of individuals has a contraindication for treatment or when guidelines already recommend that individuals above a certain probability threshold should receive treatment. This can lead to individuals receiving extreme weights, resulting in biased and imprecise estimates of model performance.[15] In addition, problems can occur due to incorrect specification of the treatment propensity model, for example due to the presence of unmeasured confounders- predictors associated with both the outcome and the use of treatment in the validation set. Variants of the basic IPW procedure can be applied, such as weight truncation, which may improve the performance of this method in settings where the assumptions are violated.[21]

Figure 2 a-d: Risk distributions in two simulated validation sets, before and after applying different approaches to correct for treatment use.

50% of individuals received an effective treatment (relative odds reduction on treatment: 0.5) (see Table 2 scenarios 2 and 1, respectively, for details). a: the model was validated on the combined treatment and control group of a randomised trial. b-d: the model was validated using data from a non-randomised setting where the probability of receiving treatment depended on an individual's (untreated) outcome risk. Solid black lines represent the observed risks in the validation set after treatment. Dashed black lines represent the risks observed after applying correction methods to the data (a-b: the exclusion of treated individuals, c: IPW, d: IPW followed by the exclusion of treated individuals). Grey lines represent the risks of the same individuals had they remained untreated.



Model recalibration

The incidence of the predicted outcome may vary between development and validation data sets. If this is the case, the predictions made by the model will not, on average, match the outcome incidence in the validation data set.[22] As discussed in section 2.1, use of an effective treatment in a validation data set will lead to fewer outcome events and thus a lower incidence than there would have been had the validation set remained untreated. One approach to account for this would be to recalibrate the original model

using the partially treated validation data set. In a logistic regression model, a derivative of the incidence of the outcome is captured by the intercept term in the model, and thus a simple solution would seem to be to re-estimate the model intercept using the validation data set.[23, 24] In doing this, the average predicted risk provided by the recalibrated model should then be equal to the (observed) overall outcome frequency in the validation set. Further details of this procedure are given in Table 1. Where treatment has been randomly allocated, intercept recalibration should indeed account for the risk-lowering effects, provided that the magnitude of the treatment effect does not vary depending on an individual's risk and thus is constant over the entire predicted probability range. In non-randomized settings, where treatment use by definition is associated with participant characteristics, a simple intercept recalibration is unlikely to be sufficient due to interactions between treatment use and patient characteristics that are predictors in the model.

However, although recalibration may seem a suitable solution for modelling the effects of treatment, when applying recalibration, concerns should also be raised over the interpretation of the estimated performance of the model. Differences in outcome incidence between the development data set and validation data set may not be entirely attributable to the effects of treatment use. By recalibrating the model to adjust for differences in treatment use and effects, we simultaneously adjust for differences in case-mix between the development and validation set. As the aim of validation is to evaluate the performance of the original prognostic model, in this case in a treatment-naïve sample, recalibration may actually lead to an optimistic impression of the accuracy of predictions made by the original model in the validation set. For example, if the validation set included individuals with a notably greater prevalence of comorbidities and thus were more likely to develop the outcome, recalibration prior to validation could mask any inadequacies of the model when making predictions in this subset of high-risk individuals. Thus recalibration is not an appropriate solution to the problem.

Incorporation of treatment in the model

A more explicit way to deal with treatment use would be to update the prognostic model with treatment use added as a new predictor. If effective, treatment can actually be considered to be a missing predictor in the original developed model. However, unlike other predictors, when validating a model in a non-randomised data set, we cannot know whether a person in practice will indeed receive the treatment at the point of making a prediction. By adding a binary predictor for treatment use to the original prognostic model, one may aim to alleviate the misfit that results from the omission of this predictor, and get closer to the actual performance of the original model in the validation set, had individuals remained untreated.

There are a number of approaches to updating a model with a new predictor.[22, 23, 25] One option would be to incorporate an indicator for treatment on top of the prognostic model, keeping the original model coefficients fixed. However, in doing this we assume that there is no correlation between treatment use and the predictors in the model. Instead the model could be entirely refitted with the addition of an indicator term for treatment using the validation data set (for further details, see Table 1). It may be necessary to include statistical interaction terms in the updated model, where anticipated.[26]

A challenge when considering this approach is the correct specification of the updated prediction model. Failure to correctly specify any interactions between treatment and other predictors in the validation set could mean that the effects of treatment are not completely taken into account. Furthermore, the addition of a term for treatment to the model that is to be validated may improve the performance beyond that of the original model due to the inclusion of additional predictive information. Thus, as with recalibration, we do not recommend this approach.

Outline of a simulation study

We assess the performance of different methods to account for the effects of treatment in fifteen scenarios using simulated data. The effectiveness of two methods described in section 2.2, model recalibration and the incorporation of a term for treatment use in the model, are not present, as their inferiority has already been discussed.

Details of the simulation study are provided in Table 2, which describes 15 scenarios that were studied. For each scenario, a development data set of 1000 individuals of whom all remained untreated throughout the study was simulated. A prognostic model was developed with two predictors using logistic regression analysis, specifying the model so it matched the data generating model. Fifteen validation sets of 1000 individuals were drawn using the same data generating mechanism as their corresponding development data sets, representing an ideal untreated validation set to estimate the model's ability to predict untreated risks. Subsequently, 50% of the individuals in each validation set were simulated to receive a risk-lowering point-treatment with a constant effect of a reduction in the outcome odds by 50%.

In scenarios 1, 3 and 4, an individual's probability of receiving treatment was a function of their untreated risk of the outcome, representing observational data. In scenario 2, treatment was randomly allocated to individuals, simulating data from a RCT. In scenarios 1 and 3, there was a moderate positive association between risk and treatment allocation, and thus individuals with a more "risky" profile were more likely to receive treatment. In scenario 4 this association was large: treatment was allocated to most (95%) of the individuals with a predicted risk higher than 18%. In scenario 3, the

relative treatment effect was allowed to increase with increasing risk. Using scenario 1 as a starting point, in scenarios 5–12, the effect of treatment on risk varied from strong to weak, and the proportion of individuals treated varied. In scenarios 13–15, an unobserved predictor with varying association (moderate negative, weak positive or strong positive) with the outcome was included in the data generating model.

The performance of the prognostic model was estimated in each of these data sets, first ignoring the effects of treatment, and again either by first excluding treated individuals from the analysis, or by applying IPW methods (as specified in Table 1). We applied standard IPW and IPW with weight truncation (at the 98th percentile). For scenarios 1–12, the treatment propensity model was correctly specified; for scenarios 13–15, the unobserved predictor was (by definition) omitted from the treatment propensity model.

Table 2: A summary of fifteen simulated scenarios.

	Data generating models (development, validation)† $logit(Y)=b_0+b_1X_1+b_2X_2+b_3U$				Treatment model $P(Tr)= logit^{-1}(a_0+a_1R)$		% treated in validation set	Relative treatment effect on Y
	b_0	b_1	b_2	b_3	a_o	a_1		
1	-1.50	1	1	0	1.95	-10	50	0.5
2	-1.50	1	1	0	0	0	50	0.5
3	-1.50	1	1	0	1.95	-10	50	$logit^{-1}(-1+5R)$
4	-1.50	1	1	0	18.0	-100	50	0.5
5	-1.50	1	1	0	3.30	-10	25	0.3
6	-1.50	1	1	0	3.30	-10	25	0.5
7	-1.50	1	1	0	3.30	-10	25	0.8
8	-1.50	1	1	0	1.95	-10	50	0.3
9	-1.50	1	1	0	1.95	-10	50	0.8
10	-1.50	1	1	0	0.70	-10	75	0.3
11	-1.50	1	1	0	0.70	-10	75	0.5
12	-1.50	1	1	0	0.70	-10	75	0.8
13	-1.55	1	1	1	1.90	-10	50	0.5
14	-1.70	1	1	2	1.80	-10	50	0.5
15	-2.15	1	1	4	1.55	-10	50	0.5

Scenario 1 is the default scenario on which all other scenarios are based. All simulated data sets had a sample size n=1000, and a 20% chance of outcome (before treatment).
Abbreviations: P(Tr): probability of treatment ; R: baseline (untreated) risk of an individual in the validation set, calculated as $logit^{-1}(Y)$, where $logit^{-1}$ is the inverse-logit function.
†Predictors X1, X2, and U were independent random draws from a normal distribution (mean = 0, variance = 0.2); the binary outcome Y was sampled from a binomial distribution with outcome probability derived from the data generating model.

In all simulated validation sets and for all methods being applied, performance was estimated in terms of the c-index (area under the ROC curve) and observed:expected (O:E) ratio. For scenarios 1-4 and 13-15 calibration plots were constructed. For IPW methods, calculated IPW weights were used to estimate weighted statistics (see Supplement 1 for further details). In order to obtain stable estimates of the c-index and O:E ratio, we repeated the process of data generation, model development and validation 10,000 times, calculating the mean and standard deviation (SD) of the distribution of the 10,000 estimates. Calibration plots were based on sets of 1 million individuals (equivalent to combining results from 1000 repeats in data sets with 1000 individuals) for each scenario. R code to reproduce the analyses can be found in Supplement 1.

Results

Results of the simulation study are presented below. A summary of the estimated performance measures in each scenario can be found in Tables 3 and 4, and calibration plots for scenarios 1-4 and 13-15 are depicted in Figures 3 and 4, respectively.

Ignore treatment

Ignoring the effects of treatment resulted, as expected, in predicted risks that were always greater than the observed outcome frequencies, suggesting poor model calibration in all scenarios. This was exacerbated in non-randomised settings, in which there appeared to be greater miscalibration in high-risk individuals. When treatment allocation was non-random, ignoring treatment led to an underestimation of the c-index by up to 0.08 (scenario 3), whereas the c-index did not noticeably change in the RCT scenario. As expected, when either the effectiveness of treatment or the proportion of individuals treated increased, both the O:E ratio and c-index were more severely underestimated.

Table 3: Estimated calibration in the validation set (observed:expected (O:E) ratio) across fifteen different simulated scenarios.

	Method					
	Reference: untreated	**Ignore treatment**	**Exclude treated**	**IPW**	**IPW, exclude**	**IPW_{trunc}, exclude**
1	1.00 (0.09)	0.76 (0.07)	1.01 (0.13)	0.79 (0.09)	1.00 (0.13)	1.00 (0.12)
2	1.00 (0.09)	0.79 (0.07)	1.00 (0.11)	0.79 (0.07)	1.00 (0.11)	1.00 (0.11)
3	1.01 (0.09)	0.69 (0.07)	1.00 (0.13)	0.76 (0.09)	1.00 (0.13)	1.00 (0.12)
4	1.00 (0.09)	0.72 (0.07)	1.01 (0.16)	0.74 (0.30)	0.98 (0.44)	1.00 (0.17)
5	1.00 (0.09)	0.80 (0.08)	1.00 (0.13)	0.68 (0.07)	1.00 (0.10)	1.00 (0.10)
6	1.00 (0.09)	0.87 (0.08)	1.01 (0.10)	0.79 (0.08)	1.00 (0.10)	1.00 (0.10)
7	1.00 (0.09)	0.96 (0.09)	1.01 (0.10)	0.93 (0.10)	1.00 (0.10)	1.00 (0.10)
8	1.00 (0.09)	0.63 (0.06)	1.01 (0.12)	0.68 (0.08)	1.00 (0.13)	1.00 (0.12)
9	1.00 (0.09)	0.91 (0.08)	1.01 (0.12)	0.92 (0.09)	1.00 (0.13)	1.00 (0.12)
10	1.00 (0.09)	0.49 (0.06)	1.00 (0.17)	0.68 (0.11)	1.00 (0.20)	1.00 (0.18)
11	1.00 (0.09)	0.66 (0.07)	1.00 (0.17)	0.79 (0.11)	1.00 (0.20)	1.00 (0.18)
12	1.01 (0.09)	0.88 (0.08)	1.01 (0.17)	0.92 (0.12)	1.00 (0.20)	1.00 (0.18)
13	1.00 (0.09)	0.75 (0.07)	0.90 (0.12)	0.76 (0.08)	0.87 (0.12)	0.88 (0.11)
14	1.00 (0.09)	0.74 (0.07)	0.70 (0.10)	0.72 (0.07)	0.67 (0.10)	0.67 (0.09)
15	1.00 (0.09)	0.76 (0.07)	0.39 (0.07)	0.74 (0.07)	0.38 (0.07)	0.38 (0.07)

Results were derived from development and validation sets of 1000 individuals. Performance estimates are the means (and standard deviations) of the distribution of O:E ratios from 10000 simulation replicates. See Table 2 for details of the scenarios.
Abbreviations: Exclude: exclusion of treated individuals from the analysis; IPW: inverse (treatment) probability weighting; IPW_{trunc}: IPW with weight truncation at 98th percentile.

Table 4: Estimated discrimination in the validation set (c-index) across fifteen different simulated scenarios.

	Method					
	Reference: untreated	**Ignore treatment**	**Exclude treated**	**IPW**	**IPW, exclude**	**IPW_{trunc}, exclude**
1	0.67 (0.02)	0.63 (0.02)	0.65 (0.03)	0.66 (0.03)	0.66 (0.05)	0.65 (0.04)
2	0.67 (0.02)	0.66 (0.02)	0.67 (0.03)	0.66 (0.02)	0.67 (0.03)	0.67 (0.03)
3	0.67 (0.02)	0.59 (0.03)	0.65 (0.03)	0.64 (0.03)	0.66 (0.05)	0.65 (0.04)
4	0.67 (0.02)	0.59 (0.03)	0.60 (0.04)	0.59 (0.08)	0.57 (0.15)	0.60 (0.05)
5	0.67 (0.02)	0.62 (0.02)	0.65 (0.03)	0.66 (0.03)	0.67 (0.03)	0.66 (0.03)
6	0.67 (0.02)	0.64 (0.02)	0.65 (0.03)	0.66 (0.03)	0.66 (0.03)	0.66 (0.03)
7	0.67 (0.02)	0.66 (0.02)	0.65 (0.03)	0.67 (0.03)	0.67 (0.03)	0.66 (0.03)
8	0.67 (0.02)	0.60 (0.03)	0.65 (0.03)	0.66 (0.03)	0.66 (0.05)	0.65 (0.04)
9	0.67 (0.02)	0.65 (0.02)	0.65 (0.03)	0.66 (0.03)	0.66 (0.05)	0.65 (0.04)
10	0.67 (0.02)	0.61 (0.03)	0.65 (0.05)	0.66 (0.05)	0.66 (0.08)	0.65 (0.06)
11	0.67 (0.02)	0.64 (0.03)	0.65 (0.05)	0.66 (0.05)	0.66 (0.08)	0.65 (0.06)
12	0.67 (0.02)	0.66 (0.02)	0.65 (0.05)	0.66 (0.04)	0.66 (0.08)	0.65 (0.06)
13	0.66 (0.02)	0.63 (0.02)	0.63 (0.03)	0.65 (0.03)	0.64 (0.05)	0.63 (0.04)
14	0.65 (0.02)	0.63 (0.02)	0.60 (0.04)	0.62 (0.03)	0.61 (0.04)	0.60 (0.04)
15	0.62 (0.02)	0.61 (0.03)	0.57 (0.05)	0.58 (0.03)	0.57 (0.05)	0.57 (0.05)

Results were derived from development and validation sets of 1000 individuals. Performance estimates are the means (and standard deviations) of the distribution of c-indexes from 10000 simulation replicates. See Table 2 for details of the scenarios.
Abbreviations: Exclude: exclusion of treated individuals from the analysis; IPW: inverse (treatment) probability weighting; IPW_{trunc}: IPW with weight truncation at 98th percentile.

Method 1: Exclude treated individuals

Excluding treated individuals resulted in calibration measures that appeared to reflect those of the untreated target population in most scenarios. However, as Figure 3 shows, use of this approach when treatment allocation is dependent on an individual's risk results in a loss of information about calibration in high risk individuals. When treatment allocation was random (scenario 2), this approach yielded a correct estimate of the c-index. As treatment allocation became increasingly associated with an individual's risk across scenarios, this method yielded lower estimates for discrimination than observed in the untreated set, due to the selective exclusion of high-risk individuals, and consequently a narrower case-mix. The estimates of the c-index and O:E ratio were constant as the treatment effect and proportion treated changed across scenarios 5–12. In the presence of a strong unmeasured predictor of the outcome associated with treatment use (scenarios 14–15), exclusion of treated individuals resulted in an underestimation of the performance of the model. In addition, in all scenarios the precision of estimates of both the O:E ratio and c-index decreased due to the reduction in effective sample size.

Method 2: Inverse probability weighting

Across all scenarios, IPW alone did not improve calibration, compared to when treatment was ignored, whereas IPW followed by the exclusion of treated individuals provided correct estimates for calibration. IPW alone or followed by the exclusion of treated individuals improved estimates of the c-index in all scenarios where the assumptions of positivity and no unobserved confounding were met. In scenario 4, where treatment allocation was determined by a strict risk-threshold and thus the assumption of positivity was violated, IPW was ineffective, and resulted in the worst estimates of discrimination across all methods. In addition, the extreme weights calculated in scenario 4 led to very large standard errors. In scenarios 13–15, the presence of an unobserved confounder led to the failure of IPW to provide correct estimates of the c-index. Weight truncation at the 98% percentile increased precision but was less effective in correcting of the c-index for the effects of treatment.

Figure 3: Calibration curves calculated in a treated validation set, following different approaches to account for the effects of treatment.

*S*cenario 1: P(treatment) increases with risk, fixed treatment effect; scenario 2: randomized treatment, fixed treatment effect; scenario 3: P(treatment) increases with risk, treatment effect increases with risk; scenario 4: 18% baseline risk threshold for treatment, fixed treatment effect. Plots were based on sets of 10^6 individuals.

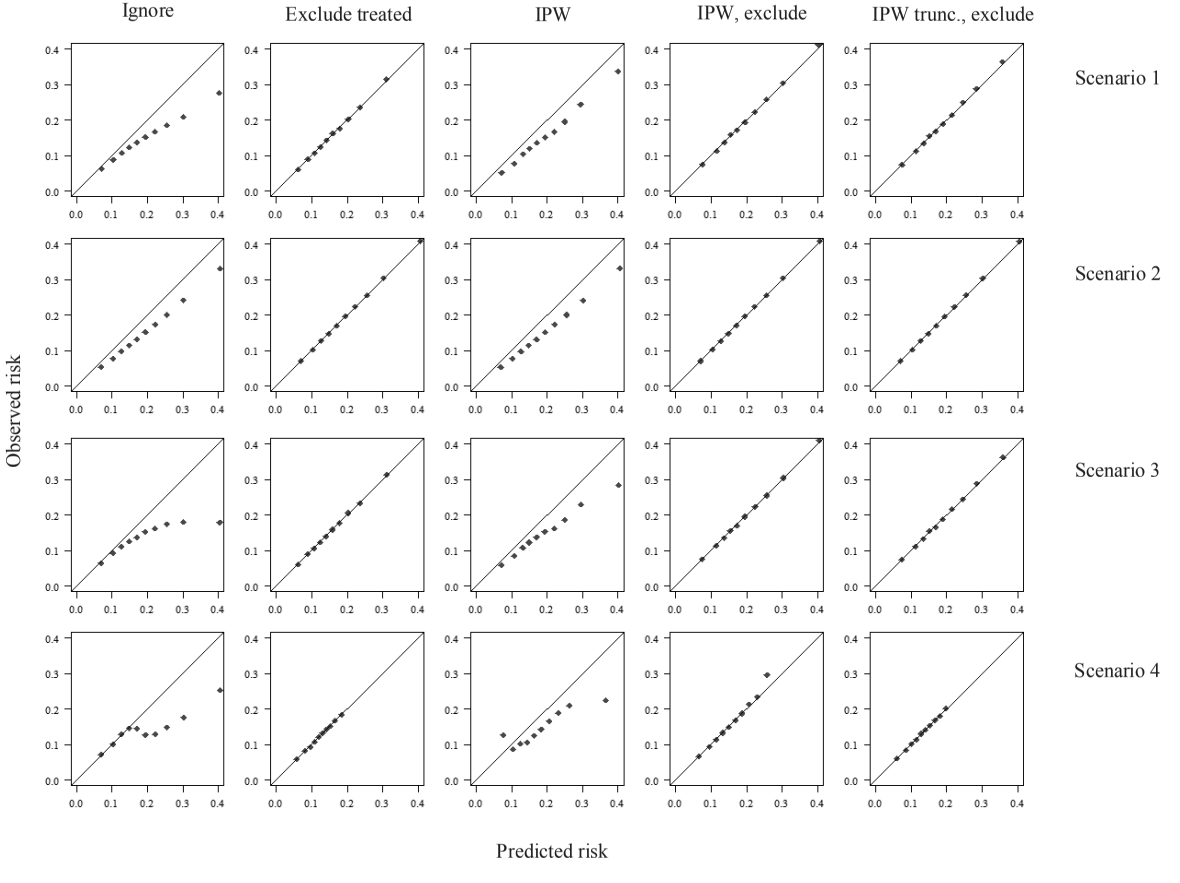
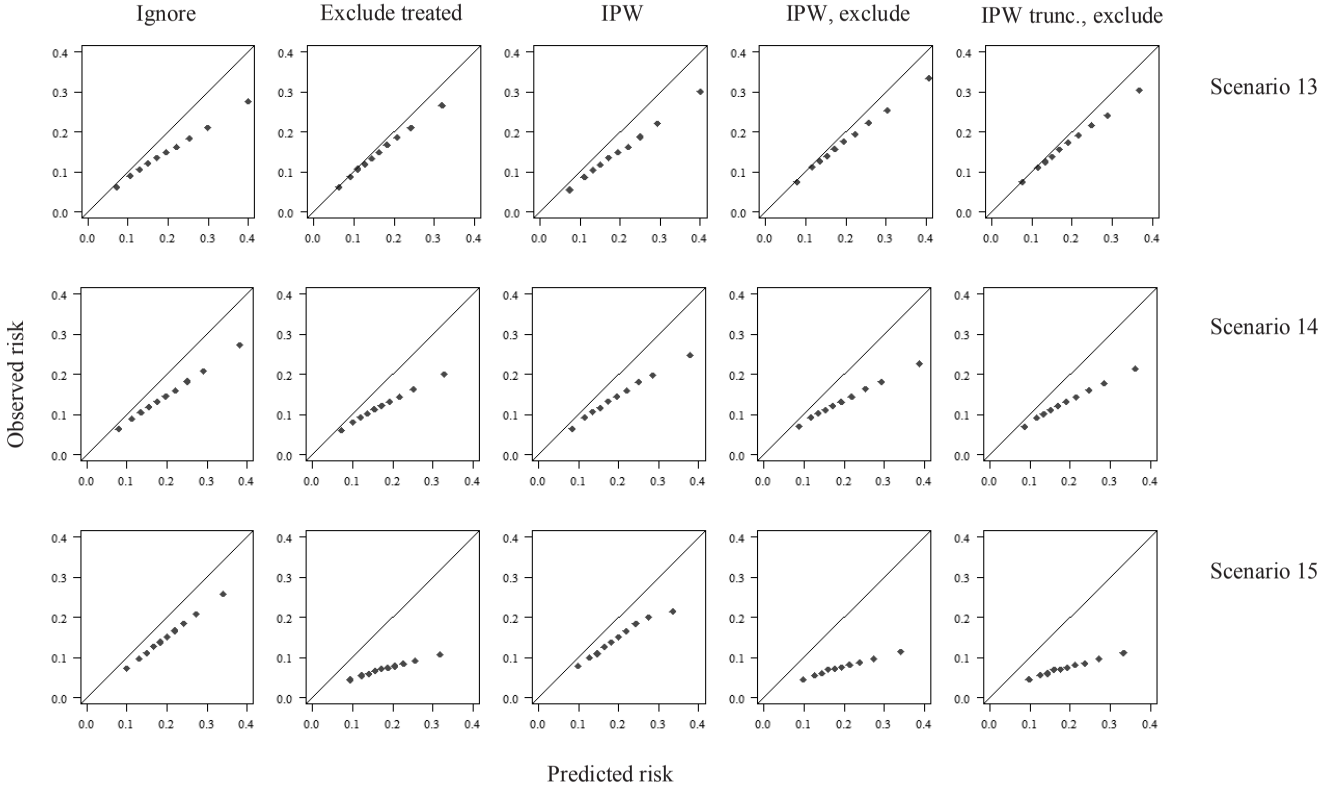


Figure 4: Calibration curves calculated in a treated validation set, following different approaches to account for the effects of treatment, in the presence of an unmeasured predictor (U) associated with both the outcome and the probability of receiving treatment.

Scenario 13: U has a weak association with the outcome (log(OR) = 1); scenario 14: U has a moderate association with the outcome (log(OR) = 2); scenario 15: U has a strong association with the outcome (log(OR) = 4). Plots were based on sets of 10^6 individuals.



Discussion

We showed that when externally validating a prognostic model that was developed for predicting "untreated" outcome risks, treatment use in the validation set may substantially impact on the performance of the model in that validation set. Treatment use is problematic, if ignored, regardless of how treatment has been allocated, though more challenging to circumvent when non-randomized. While the risk-lowering effect of treatment seems to have little effect on model discrimination in randomised trial data, the model will appear to systematically over-estimate risks (miscalibration). This effect worsens with greater dependency of treatment use on patient characteristics (e.g. baseline risk).

We present simple methods that could be considered when attempting to take the effects of treatment use into account. While the use of IPW in prediction model research is uncommon, the rationale behind using IPW in settings with non-randomized treatments is motivated by its use to remove the influence of treatment on causal (risk) factor-outcome associations.[27, 28] Although the use of IPW prior to the exclusion of treated individuals is a promising solution in data where treatments are non-randomly allocated, it should not be used when there are severe violations of the underlying assumptions, e.g. in the presence of non-positivity (where some individuals had no chance of receiving treatment), or when there is an unobserved confounder, strongly associated with both the outcome and treatment use. There is thus a need to explore alternative methods to IPW to account for the effects of treatment use when validating a prognostic model in settings with non-random treatment use.

Although the results of our simulations support the expected behaviour of the methods described in section 2.2, some findings warrant further discussion. First, although excluding treated individuals when treatment use is non-random theoretically results in incorrect estimates of model performance, in our simulations, the impact on model discrimination was small in most scenarios. However, when the association between an individual's risk profile and the chance of being treated increased (scenario 4), the selection bias due to excluding treated individuals resulted in a large decrease in the c-index, as expected. Second, in simulated scenarios in which an unobserved confounder of the treatment-outcome relation was present, the performance of the model greatly decreased after excluding treated individuals, with or without IPW. This is likely due to the selective exclusion of individuals with a high value for the strongly predictive unobserved variable. This results in a narrower case-mix distribution, and consequently lower model discrimination, as well as miscalibration due to the exclusion of a strong predictor of the outcome.

While it is unclear to what extent treatment use has affected existing prognostic model validation studies, findings from a systematic review of cardiovascular prognostic model studies indicate that changes in treatment use after baseline measurements in a validation study are rarely considered in the analysis.[29] While a number of studies excluded prevalent treatment users from their analyses, the initiation of risk-lowering interventions, such as statins, revascularization procedures and lifestyle modifications during follow-up was not taken into account. An equally alarming finding was that very few validation studies even reported information about treatment use during follow-up, raising concerns over the interpretation of the findings of these studies. Based on the findings of the present study, we suggest that information about the use of effective treatments both at the study baseline and during follow-up should be reported in future studies.

It must be noted that not all prediction model validation studies require the same considerations for treatment use. Although we have discussed prognostic models used for predicting the risk of an outcome without treatment, sometimes prognostic models are developed for making predictions in both treated and untreated individuals. If, for example, the treatments used in the validation set are a part of usual care, and are present in the target population for the model, then differences in the use of these treatments between the development and validation sets should be viewed as a difference in case-mix and not as an issue that we need to remove. Furthermore, if the model adequately incorporates relevant treatments (e.g. through the explicit modelling of treatment use), differences in treatment use between the development and validation sets can again be viewed as a difference in case-mix. In the event that treatments have not been modelled (e.g. because a new treatment has become readily available since the development of the model), the model could be updated through recalibration, or better yet by including a term for treatment in the updated model, leading to a completely new model, which in turn would require validation. Researchers must therefore first identify which treatments used in a validation data set could bias estimates of model performance, if ignored.

There are limitations to the guidance that we provide. First, we do not present a complete evaluation of all possible methods across a range of different settings, which would require at least an extensive simulation study. We argue, however, that the logical argumentation provided for each method forms a good starting point for further investigation. Furthermore, the list of methods that we present is by no means exhaustive and we encourage the consideration and development of new approaches for more complex settings, such as time-to-event settings, and where limited sample sizes pose a challenge. Second, we assumed for simplicity that a model has been developed in an untreated data set. In reality, it is likely that a model has been developed also in a partially treated set. The considerations for validation then remain the same, but

it should be noted that failure to properly account for the effects of treatment in the development of a model can lead to a model that underestimates untreated risks.[13] Third, for simplicity we considered single point treatments in our simulated examples. Patterns of treatment use in reality are often complex, with individuals receiving multiple non-randomized treatments, even in RCTs. Finally, we also recognize that while this paper discusses the validation of prognostic models, the same considerations for treatment use can, in some circumstances, be relevant to diagnostic studies (i.e. where treatment between index testing and outcome verification could lead to similar- and even more serious- problems).

Conclusion

When validating a previously developed prediction model for predicting risks without treatment in another data set, failure to properly account for (effective) treatment use in that validation sample will likely lead to poor performance of the prediction model and thus measures should be taken to remove the effects of treatment use. When validating a model with data in which treatments have been randomly allocated, simply excluding treated individuals is sufficient, at the cost of a loss of precision. In observational studies, where treatment allocation depends on patient characteristics or risk, inverse probability weighting followed by the exclusion of treated individuals can provide correct estimates of the actual performance of the model in its target population.

References

1. Moons, K.G., et al., Risk prediction models: II. External validation, model updating, and impact assessment. Heart, 2012. **98**(9): p. 691-8.
2. Moons, K.G., et al., Prognosis and prognostic research: what, why, and how? Bmj, 2009. **338**: p. b375.
3. Steyerberg, E.W., et al., Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med, 2013. **10**(2): p. e1001381.
4. Collins, G.S., et al., External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol, 2014. **14**: p. 40.
5. Harrell, F.E., Jr., et al., Regression modelling strategies for improved prognostic prediction. Stat Med, 1984. **3**(2): p. 143-52.
6. Harrell, F.E., Jr., K.L. Lee, and D.B. Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med, 1996. **15**(4): p. 361-87.
7. Justice, A.C., K.E. Covinsky, and J.A. Berlin, Assessing the generalizability of prognostic information. Ann Intern Med, 1999. **130**(6): p. 515-24.
8. Debray, T.P., et al., A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol, 2015. **68**(3): p. 279-89.
9. Vergouwe, Y., K.G. Moons, and E.W. Steyerberg, External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol, 2010. **172**(8): p. 971-80.
10. Riley, R.D., 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**: p. i3140.
11. Liew, S.M., J. Doust, and P. Glasziou, Cardiovascular risk scores do not account for the effect of treatment: a review. Heart, 2011. **97**(9): p. 689-697.
12. Muntner, P., et al., Comment on the reports of over-estimation of ASCVD risk using the 2013 AHA/ACC risk equation. Circulation, 2014. **129**(2): p. 266-7.
13. Groenwold, R.H., et al., Explicit inclusion of treatment in prognostic modelling was recommended in observational and randomised settings. J Clin Epidemiol, 2016.
14. Robins, J.M., M.A. Hernan, and B. Brumback, Marginal structural models and causal inference in epidemiology. Epidemiology, 2000. **11**(5): p. 550-60.
15. Lee, B.K., J. Lessler, and E.A. Stuart, Weight trimming and propensity score weighting. PLoS One, 2011. **6**(3): p. e18174.
16. Cole, S.R. and M.A. Hernán, Constructing Inverse Probability Weights for Marginal Structural Models. American Journal of Epidemiology, 2008. **168**(6): p. 656-664.
17. Austin, P.C. and E.A. Stuart, Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med, 2015. **34**(28): p. 3661-79.
18. Janssen, K.J., et al., Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol, 2008. **61**(1): p. 76-86.

19. Steyerberg, E., Clinical prediction models: a practical approach to development, validation, and updating. 2008: Springer Science & Business Media.
20. Austin, P.C. and E.W. Steyerberg, 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**: p. 82.
21. Rosenbaum, P.R. and D.B. Rubin, The central role of the propensity score in observational studies for causal effects. Biometrika, 1983. **70**(1): p. 41-55.
22. Pfeffermann, D., The role of sampling weights when modeling survey data. International Statistical Review/Revue Internationale de Statistique, 1993: p. 317-337.
23. Petersen, M.L., et al., Diagnosing and responding to violations in the positivity assumption. Stat Methods Med Res, 2012. **21**(1): p. 31-54.
24. van Houwelingen, H.C., Validation, calibration, revision and combination of prognostic survival models. Stat Med, 2000. **19**(24): p. 3401-15.
25. Su, T.-L., et al., A review of statistical updating methods for clinical prediction models. Statistical Methods in Medical Research, 2015.
26. van Klaveren, D., et al., Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. J Clin Epidemiol, 2015. **68**(11): p. 1366-74.
27. Wang, Y. and Y. Fang, Adjusting for treatment effect when estimating or testing genetic effect is of main interest. Journal of Data Science, 2011. **9**(1): p. 127-138.
28. Spieker, A.J., J.A. Delaney, and R.L. McClelland, Evaluating the treatment effects model for estimation of cross-sectional associations between risk factors and cardiovascular biomarkers influenced by medication use. Pharmacoepidemiol Drug Saf, 2015. **24**(12): p. 1286-96.
29. 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 Dec;1(1):15.

Supplemental material

Supplement 1: R code to implement methods and replicate the simulation study

```
# (R version 3.3.2 (2016-10-31))
# Code can be run to reproduce Table 3 and Table 4. Calibration plots can be
# reproduced by setting n = 1e6, n.sim = 1 and appraisal(..., plots=TRUE).
# Different scenarios can be run by removing # symbols where appropriate.

require(rms)
require(boot)

# Function to return discrimination and calibration
appraisal <- function(obs, LP, N, plots = TRUE, title = "", w=rep(1,length(obs))){
  sum.tab<- matrix(rep(NA, 2), nrow=1, dimnames=list(NULL, c("O/E ratio", "C
index")))
 pred <- 1 / (1 + exp(-LP))
 sum.tab[1] <- (sum(obs * w)/ sum(w)) / (sum(pred * w) / sum(w)) # O:E ratio
 sum.tab[2] <- somers2(LP,obs, weights = w)[1] #c-index
 if(plots){ #calibration plot
  predw <- rep(pred, round(w))
  obsw <- rep(obs, round(w))
  mf<- data.frame(cbind(obsw, predw))
  mf<- mf[with(mf, order(predw)),]
  dec.pred <- dec.obs <- c()
  for (i in 1:10) { # create deciles for calibration plot
   dec.obs[i] <- mean(mf[(1+(i-1)*(sum(round(w))/10)):((sum(round(w))/10)*i), 1])
   dec.pred[i] <- mean(mf[(1+(i-1)*(sum(round(w))/10)):((sum(round(w))/10)*i), 2])
  }
  plot(dec.pred,dec.obs, pch = 18, cex=1.5, main= paste(title),
      col = "darkgreen", bg= "darkgreen", xlim = c(0,0.4), ylim = c(0,0.4),
      ylab = "Observed risk", xlab = "Predicted risk", cex.lab = 1.25, las=1)
  abline(a=0,b=1)
 }
 return(round(sum.tab,4))
}

# Set up simulation
n = 1000
n.sim = 10000
mu = 0
```

```
s = sqrt(0.2)
b0U <- -1.5 # For scenarios 14-16, choose: (-1.55, -1.7, -2.15)
bU <- 0 # For scenarios 14-16, choose: (1, 2, 4)
Z <- -1.95 # For scenarios 5-13, choose: (-3.3 (25% treated),-1.95 (50%), -0.7 (75%))
# For scenarios 14-16, choose: (-1.9, -1.8, -1.55)
t.eff <- 0.5 # For scenarios 5-13: (0.3, 0.5, 0.8)
untreated <- ignore <- restrict <- ipw <- ipw.r <- ipw.tr <- matrix(NA,n.sim, ncol=2)

# Simulation
for(i in 1:n.sim) {
# Scenarios 1-12
# Generate development set (untreated outcomes)
 df1 <- data.frame(X1=rnorm(n, mu, s), X2=rnorm(n, mu, s))
 drisk <- 1/(1+exp(-(-1.5 + 1*df1$X1 + 1*df1$X2)))
 dY <- rbinom(n, 1, drisk)
# Generate validation set (untreated outcomes)
 df2 <- data.frame(X1=rnorm(n, mu, s), X2=rnorm(n, mu, s))
 vrisk <- 1/(1+exp(-(-1.5 + 1*df2$X1 + 1*df2$X2)))
 vY <- rbinom(n, 1, vrisk)

# Scenarios 13-15
# Generate development set (untreated outcomes)
#  df1 <- data.frame(X1=rnorm(n, mu, s), X2=rnorm(n, mu, s), U=rnorm(n, mu, s))
#  drisk <- 1/(1+exp(-(b0U + 1*df1$X1 + 1*df1$X2 + bU*df1$U)))
#  dY <- rbinom(n, 1, drisk)
# Generate validation set (untreated outcomes)
#  df2 <- data.frame(X1=rnorm(n, mu, s), X2=rnorm(n, mu, s), U=rnorm(n, mu, s))
#  vrisk <- 1/(1+exp(-(b0U + 1*df2$X1 + 1*df2$X2 + bU*df2$U)))
#  vY <- rbinom(n, 1, vrisk)

# Generate treated outcomes
# Scenario 1, 5-15
 pt   <- 1/(1+exp(-(Z + 10*vrisk)))
 treated <- rbinom(n,1,pt)
 lodds <- log(vrisk/(1-vrisk))
 lodds[treated == 1] <- lodds[treated == 1] + log(t.eff)
 lodds <- inv.logit(lodds)
 Y.treated <- rbinom(n, 1, lodds)
# Scenario 2
#   treated          <- rbinom(n, 1, 0.5)
```

```
#   lodds <- log(vrisk/(1 - vrisk))
#   lodds[treated == 1] <- lodds[treated == 1] + log(0.5)
#   lodds <- inv.logit(lodds)
#   Y.treated <- rbinom(n, 1, lodds)
# Scenario 3
#  t.OR <-1 / (1+exp(-(1-5*vrisk)))
#  pt <- 1/(1+exp(-(Z + 10*vrisk)))
#  treated          <- rbinom(n,1,pt)
#  lodds <- log(vrisk/(1-vrisk))
#  lodds[treated == 1] <- lodds[treated == 1] + log(t.OR)[treated==1]
#  lodds <- inv.logit(lodds)
#  Y.treated <- rbinom(n, 1, lodds)
# Scenario 4
#  pt   <- 1/(1+exp(-(-18 + 100*vrisk)))
#  treated  <- rbinom(n,1,pt)
#  lodds <- log(vrisk/(1-vrisk))
#  lodds[treated == 1] <- lodds[treated == 1] + log(0.5)
#  lodds <- inv.logit(lodds)
#  Y.treated <- rbinom(n, 1, lodds)

# Develop model (untreated)
# For Scenarios 13-15, unobserved predictor U is not included in the prediction model
  model1 <- glm(dY ~ df1$X1 + df1$X2, family=binomial)
  vLP <- model1$coef[1] + model1$coef[2]*df2$X1 + model1$coef[3]*df2$X2
  vpredrisk <- 1/(1+exp(-(vLP)))

# Inverse probability weighting
# For Scenarios 13-15, unobserved predictor U is not included in the propensity model
  psm<-glm(treated~vpredrisk, family="binomial")$fitted.values
  df2$ps <- 0
  df2[treated==0,]$ps <- 1/(1 - psm[treated==0])
  df2[treated==1,]$ps <- 1/psm[treated==1]
# Truncation (98%, upper end of weight distribution truncated)
  pstrunc<-df2[treated==0,]$ps
  pstrunc[pstrunc > quantile(pstrunc, 0.98)] <- quantile(pstrunc, 0.98)

# Assess model performance following each analytical approach
#  For "restrict" and "ipw.r", only the untreated subset is included.
#  For "ipw", "ipw.r" and "ipw.tr", performance measures are weighted.
  untreated[i,] <- appraisal(vY, vLP, n, plots=F)
```

```
  ignore[i,] <- appraisal(Y.treated, vLP, n, plots=F)
 restrict[i,] <-appraisal(Y.treated[treated==0], vLP[treated==0], length(vLP[treated==0]),
plots=F)
  ipw[i,] <- appraisal(Y.treated, vLP, n, plots=F, w=df2$ps)
  ipw.r[i,] <- appraisal(Y.treated[treated==0], vLP[treated==0],
                         length(vY[treated==0]), plots=F, w=df2[treated==0,]$ps)
  ipw.tr[i,] <- appraisal(Y.treated[treated==0], vLP[treated==0],
                           length(vY[treated==0]), plots=F, w=pstrunc)
}

# Generate output tables
full.output <- cbind(untreated, ignore, restrict, ipw, ipw.r, ipw.tr)
output.est<-matrix(apply(full.output, 2, mean),ncol=2, byrow=T)
output.SE <- matrix(apply(full.output, 2, sd), ncol=2, byrow=T)
output <- cbind(output.est[,1],output.SE[,1],output.est[,2],output.SE[,2])
```

CHAPTER 5

Accounting for time-varying treatment use when developing a prognostic model from observational data: A comparison of approaches

Abstract

Background

Failure to account for time-varying treatment use when developing a model to predict untreated risks can result in biased future predictions.

Methods

We compared approaches to develop a prognostic model for 5-year mortality risk without selective β-blocker (SBB) treatment using data from 1906 patients; 325 received SBBs during follow-up. Seven Cox regression modelling strategies were compared: 1) ignoring SBB treatment, 2) excluding SBB users or 3) censoring them when treated, 4) inverse probability of treatment weighting after censoring, including SBB treatment as a 5) binary or 6) time-varying covariate, and 7) marginal structural modelling.

Results

Compared to (1), approaches (2) and (5) provided predictions that were 1% and 2% higher on average. Performance (c-statistic, Brier score, calibration slope) varied minimally between approaches.

Conclusion

Although ignoring treatment is theoretically inferior, differences between approaches in our case study were modest. Further case studies and simulation studies should investigate when certain approaches are preferred.

Background

Prognostic models provide information that can be used to guide physicians and patients when making medical decisions. For example, if a patient has a high predicted probability of a poor health outcome, the physician may recommend preventative medication or referral for specialist treatment. For a prognostic model to be useful for guiding individual decisions regarding interventions, predictions provided by the model should ideally reflect the risk of an individual developing a certain outcome if there were to be no intervention.[1] This has been termed the "untreated risk" of an outcome [2, 3] and can be formally defined as the probability $Pr(Y^{T=0} = 1 \mid P)$, where $Y^{T=0}$ is the outcome if individuals were to remain untreated, and P is a vector of predictors.[4]

Data used to develop prognostic models often come from observational studies or longitudinal medical databases (e.g. electronic health record data) in which the study sample comprises of both treated and untreated individuals. In particular, individuals who were not receiving treatment when entering the cohort may start treatment during follow-up - an issue termed treatment "drop-in".[1,4] Unfortunately, conventional approaches to develop a prognostic model using data from a partly treated study population will result in a model that does not provide predictions of untreated risk, instead predicting $Pr(Y = 1 \mid P)$. Assuming the treatment is effective in reducing the risk of developing the outcome of interest, the observed risk of individuals in the development sample will, on average, be lower than the risk had they remained untreated. Therefore, risk predictions for future patients provided by the model will underestimate the true untreated risk of the outcome.[2] This could have a number of negative consequences, including possible under-treatment in future patients. Given that a contemporary yet untreated population in many cases no longer exists, several researchers have suggested that the effects of treatment should be accounted for in the analysis when developing a prognostic model.[1-7] However, as of yet there is no clear consensus over whether one analytical approach to deal with the effect of treatment is to be preferred over another.

Simulation studies have been conducted to investigate methods to account for the effects of treatments on prognostic model predictions,[2,4] but the effect of using different methods has, to our knowledge, not yet been investigated in real data. We aim to demonstrate, in an example from clinical practice, approaches that can be used to account for time-varying treatment drop-in in a development cohort and investigate whether the choice of approach affects a prognostic model in terms of model parameters and prognostic performance. We will first introduce a clinical case study and further explain the problem. We will then describe the approaches we compare to deal with treatment drop-in, followed by the results of this comparison.

Methods

2.1 Case study

In our case study we aim to develop a prognostic model to predict 5-year mortality risk for patients with chronic obstructive pulmonary disease (COPD) if they were not to use cardio-selective β-blocking agents (SBBs). However, in practice some COPD patients do use SBBs and these agents are known to improve survival in patients with COPD;[8,9] all other "background" treatments are considered to be a part of routine care and should not bias the estimation of untreated risk.[3] A summary of the case study is presented in Supplemental figure 1.

2.1.1 Data

A cohort was defined using electronic health record data from the Utrecht General Practitioners Network (HNU) database, with patient entry and follow-up from 1st July 1996 to 31st December 2005.[10,11] Prevalent and incident COPD cases entered the cohort on 1st July 1996 and on their date of diagnosis, respectively. Patients younger than 45 years or using β-blocker medication (ATC code C07) within 14 days prior to cohort entry were excluded. In total, 1906 patients were included; median follow-up was 6.7 years and 559 (29.3%) patients died during follow-up, of which 372 died within 5 years.

Patients prescribed SBBs during follow-up according to their individual prescription records [11] were classified as SBB "users" (yes/no). Prescription duration (days) was calculated as the total number of tablets prescribed divided by the daily tablet quantity prescribed. When prescriptions overlapped, the total number of overlapping days were added to the final prescription. In total, 325 (17.1%) patients began using SBBs after entering the cohort and the median time on treatment was 10.3 months. Prescriptions varied considerably from patient to patient (Supplemental figure 2). In total, 68 SBB-users (20.1%) died during follow-up compared to 491 non-users (31.1%) and SBB-users more often had a history of cardiovascular co-morbidities than non-SBB users (Table 1). For all other treatments, only data on use at baseline were recorded.

Missing data on the number of tablets prescribed (1.2%), and smoking status (33.3%) were singly imputed using predictive mean matching and polytomous logistic regression (mice package version 2.25).[12]

Table 1: Baseline characteristics of 1906 patients with a diagnosis of COPD stratified by selective b-blocker use during follow-up.

Characteristics	All Patients n = 1906	Selective β-blocker use n = 325	No selective β-blocker use n = 1581	p-value of difference between groups
Age at study entry *	67.8 (10.9)	67.1 (10.4)	68.0 (11.0)	0.20
Male sex	1005 (52.7)	168 (51.7)	837 (52.9)	0.73
Smoking †				0.05
Current smokers	708 (37.1)	104 (32.0)	604 (38.2)	
Former smokers	934 (49.0)	179 (55.1)	755 (47.8)	
Never smokers	264 (13.9)	42 (12.9)	222 (14.0)	
Hypertension	716 (37.6)	228 (70.2)	488 (30.9)	< 0.005
Diabetes	312 (16.4)	86 (26.5)	226 (14.3)	< 0.005
History of cardiovascular diseases				
Angina pectoris	302 (15.8)	121 (37.2)	181 (11.4)	< 0.005
Myocardial infarction	86 (4.5)	34 (10.5)	52 (3.3)	< 0.005
Atrial fibrillation	183 (9.6)	51 (15.7)	132 (8.3)	< 0.005
Heart failure	475 (24.9)	99 (30.4)	376 (23.8)	0.01
Stroke	139 (7.3)	35 (10.7)	104 (6.6)	0.01
Peripheral arterial disease	134 (7.0)	40 (12.3)	94 (5.9)	< 0.005
Drug use (ATC code) at baseline				
Drugs for obstructive airway diseases (R03)	432 (22.7)	61 (18.8)	371 (23.5)	0.08
Corticosteroids for systemic use (H02)	98 (5.1)	15 (4.6)	83 (5.2)	0.74
Lipid modifying agents (C10)	79 (4.1)	12 (3.7)	67 (4.2)	0.77
Agents acting on the renin-angiotensin system (C09)	159 (8.3)	38 (11.7)	121 (7.7)	0.02
Diuretics (C03)	236 (12.4)	49 (15.1)	187 (11.8)	0.13
Antithrombotic agents (B01)	135 (7.1)	24 (7.4)	111 (7.0)	0.91

Absolute numbers and percentages are reported. Characteristics of users and non-users were compared using t-tests (age) and chi-squared tests. * age in years; mean (standard deviation) reported. † Missing values were singly imputed (33.3%).

2.1.2 Derivation of a hypothetical prognostic model

First, a prognostic model that completely ignored SBB drop-in was developed using Cox regression using the full follow-up time of all patients. Candidate predictors were selected based on literature,[13,14] the availability of measurements, and backwards selection using likelihood ratio testing. The prediction model consisted of the estimated 5-year baseline hazard and seventeen predictors: age, sex, smoking status, history of comorbidities (diabetes, hypertension, angina pectoris, myocardial infarction, atrial fibrillation, heart failure, stroke and peripheral arterial disease) and baseline drug use (drugs for obstructive airway diseases, corticosteroids for systemic use, lipid modifying agents, agents acting on the renin-angiotensin system, diuretics and antithrombotic agents). No serious violations of the proportional hazards assumption were found. As treatment use is typically ignored in most prognostic model development studies, this model- now referred to as "model 1"- served as a reference to compare different approaches to account for treatment drop-in (section below).

2.2 Approaches to account for treatment use

We consider six other approaches that try to remove the effect of SBB drop-in in the development data. The first three approaches (models 2-4) remove treatment from the dataset in different ways. The next two approaches (models 5-6) explicitly model treatment effects. Finally, we apply a marginal structural modelling approach (model 7). All approaches are based on developing a Cox model with the same predictors as in model 1 (except when treatment is added to the model). In this section, we describe the effect of each approach on the associations between predictors in our model (P), SBB use (T), and mortality (Y), using causal diagrams. Figure 1a represents these relationships when treatment is ignored.

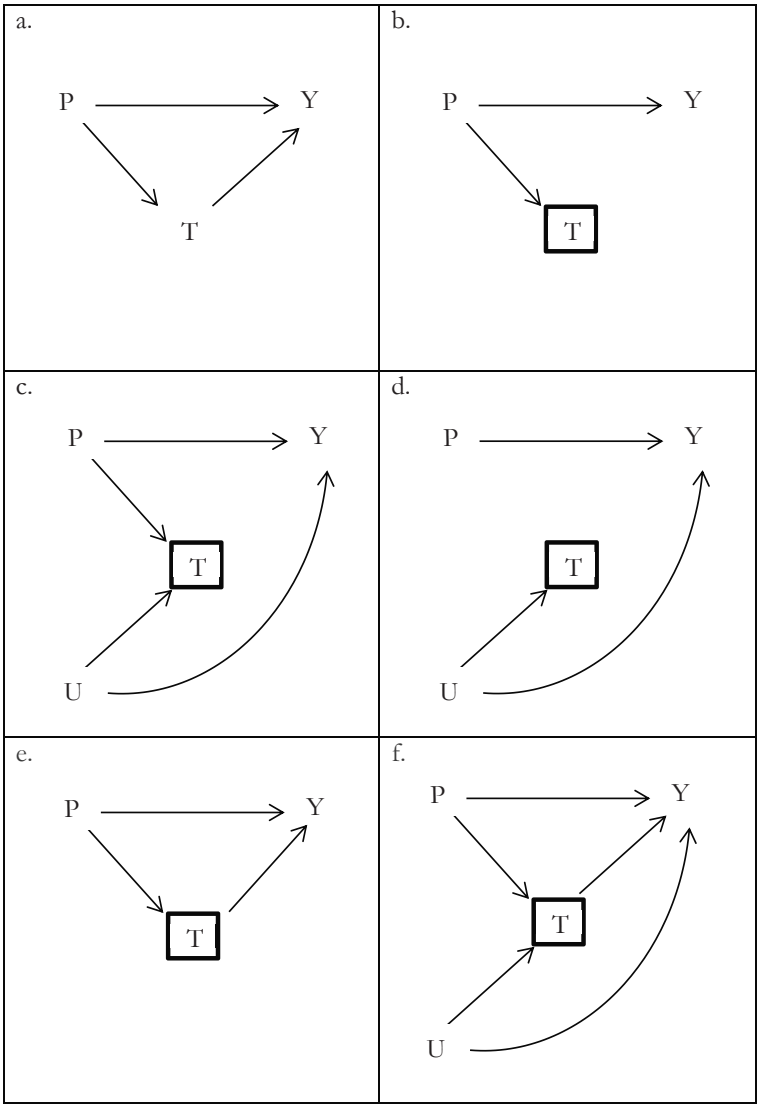
2.2.1 Excluding or censoring treatment users from the cohort

A seemingly straightforward way to remove the effect of treatment drop-in is to restrict the analysis to patients who do not receive the treatment during follow-up (restriction, model 2) or to include follow-up of those patients until the moment they start treatment and censor them from that moment onwards (censoring, model 3).

As illustrated by Figure 1b, restricting the analysis to individuals who did not receive an SBB prescription (T = 0) will remove the effect of SBBs on mortality from the cohort. However, as previously mentioned, in observational data, treatment is typically associated with predictors in the model (P), as in our example (see Table 1). Restricting the analysis to the T = 0 subset has two problems: i) the distribution of predictors and outcomes in the subset may not represent the target population for the model, as in our example where SBB users are more severely ill (Table 1); ii) in the presence of an

Figure 1: The mechanisms underlying approaches to remove the effect of SBB drop-in on a prognostic model.

For a model to provide predictions of the risk of mortality if an individual were not to receive SBBs the model parameters should be estimated such that the associations between P and Y are not influenced by the effect of T on Y. P: predictor(s), Y: outcome, T: treatment, U: variable(s). See text for further details.



unobserved confounder (U) of the treatment-outcome association, T now serves as a collider, which is represented by the two incoming arrows. Conditioning on T will alter the relation between P and U which may lead to biased estimation of the prognostic model (the association between P and Y) (see Figure 1c) (15). Alternatively, censoring individuals at the moment of SBB drop-in will retain a representative risk distribution in the data at baseline and retain more information on follow-up. However, as with restriction, if censoring is "informative", i.e. associated with P and U, the estimated model coefficients will be biased, along with predictions of $Pr(Y^{T=0} = 1 \mid P)$.

2.2.2 Reweighting using inverse probability of (censoring) treatment (IPTW) weights

To solve the issue of informative censoring, weights can be applied to patients so that the association between treatment and predictors is removed. Given that treatment drop-in does not occur at a fixed point in time, time-varying weights need to be derived.[16-18] To account for informative censoring of SBB drop-in in our case study we divided the data into seven time periods of approximately one year and derived stabilized weights for each time-period using logistic regression.[16] The prognostic model is then developed, after censoring, using a weighted Cox model (model 4). As Figure 1d indicates, if the underlying assumptions of the weighting procedure [17] are met, the association between P and T should be removed, allowing for correct estimation of the associations between P and Y.

2.2.3 Explicitly modelling treatment

Previous findings have suggested that the explicit modelling of treatment can correct for the effect of treatment on predictor-outcome associations.[2] By conditioning the prognostic model on treatment use, the pathway from P to Y via T is blocked (Figure 1e). A simple approach is to include an indicator variable for treatment drop-in in the prognostic model (model 5 in our comparison); to make prognostic predictions, all patients would then have their value for "future treatment" set to zero. However, drop-in occurs by definition after baseline and conditioning on future treatment (and thus future survival) introduces immortal time bias into the estimated treatment effect.[19] As T is associated with P, this bias will affect the coefficients of the prognostic model, and hence the predicted probabilities. To address this bias, treatment can be included as a time-varying covariate in the prognostic model.[20] Unlike all the previous approaches, this approach directly models all changes in treatment, such as discontinuation of treatment over time. However, as Figure 1f shows, residual confounding of the association between T and Y opens an additional path from P to Y (via U) and may result in model coefficients that are still biased.

2.2.4 Marginal structural modelling

This approach formally derives a prognostic model within a counterfactual framework. As described elsewhere,[4] to estimate the counterfactual, untreated risk $Pr(Y^{T=0} = 1 \mid P)$, we can first estimate stabilized IPTW weights (as described in section 2.2.2). A weighted prognostic model is derived, which includes the relevant predictor variables, as well as a term to model time-varying treatment, and terms for any interactions between treatment and other predictors in the model that may vary over time. As with explicitly modelling treatment, this approach aims to remove the association between treatment and predictors in the model, but it is also sensitive to unmeasured confounding of the treatment-outcome association, as shown in Figure 1f.

2.3 Comparison of approaches

All seven approaches were compared in terms of the resulting model coefficients, risk predictions provided by the model for patients in the development data, and the model's predictive performance. Predicted risks of all models were compared to the predicted risks from model 1. Predictive performance (Harrell's c-index, Brier score, calibration slopes) was assessed in two data sets derived from the development data: 1) the full data with patients censored at the moment of SBB treatment, and 2) SBB non-users only, to better evaluate performance in an untreated population than the full data.[21] All analyses were performed using the R statistical programme, version 3.2.2.

In addition, comparisons were repeated using a "highly-treated" subset of the data (50% of the patients began using SBBs), to see whether the results of the modelling approaches differed more with more treatment use in the development sample.

Results

The regression coefficients differed between all seven models (Table 2). The largest changes in regression coefficients compared to model 1- developed by the conventional approach of ignoring treatment - were in models 2 (SBB users excluded), 3 (SBB users censored), 4 (IPTW) and 7 (MSM). Model 6 (time-varying SBB) hardly deviated from model 1. The regression coefficient for (binary) treatment use was negative (-0.56), but became positive when modelled directly as a time varying covariate (0.39) or in a MSM (0.06).

Risk predictions made by all models in the development data ranged across the whole spectrum of probabilities (0-1). As shown by Figure 2, models 2 and 5 produced slightly higher risk predictions than model 1 (treatment ignored). Models 3 and 4 and 6 provided slightly lower risk predictions than model 1. The means of the predicted risks were 21.3% (model 1), 22.7% (model 2), 20.8% (model 3), 20.8% (model 5), 22.5%

(model 5), 21.1% (model 6) and 21.6% (model 7). When evaluated in the full data, after censoring SBB-users, predictive performance was consistent across the models: c-statistic = 0.79, Brier score = 0.09, calibration slope ranged from 0.99 (model 4) to 1.02 (Model 1,5,6). Similar results were seen when the models were evaluated in the data with SBB-users excluded.

Repeating the model development and evaluation in a highly-treated subset of the development data intensified the differences between the approaches. As Figure 3 shows, the trends of higher predictions as compared to model 1 (mean predicted risk: 18.6%) after excluding treatment (23.8%) and modelling binary treatment (23.5%), and lower predictions after censoring with (15.7%) or without (15.9%) IPTW, were more prominent.

Discussion

This study presents seven approaches to account for treatment use when developing a prognostic model using observational data. Previous work has focussed on methods to account for point treatments, i.e. treatment use that does not change over time.[2,21] As seen in our case study (see Supplemental figure 2), the use of treatments is often not fixed over time but can vary greatly in time and across individuals. The seven models, derived by ignoring treatment or applying one of the seven methods, varied in their coefficients and estimated slightly different risks for individuals in the development data set. We expected the use of effective treatments in the development set to result in a model that underestimates untreated risks, and therefore that approaches to correct for this would result in models that yielded higher predicted risks, on average. However, in our case study the impact was minimal.

There are a number of explanations for the limited, and where present, unexpected, differences between approaches that we observed in this case study. First, it may be that the effect of treatment in this development cohort was too small to impact on the performance of a prognostic model. The abundant literature supporting the protective effect of SBBs on mortality for patients with COPD (suggesting a risk ratio of ~0.7) suggests there was indeed a true effect of treatment in the patients in our dataset who did receive treatment. However, it may be that the 17.1% drop-in rate was insufficient to affect predictions. Additionally, most patients were prescribed SBBs for less than one year, which may have limited the impact on their overall risk profile during the study. With the exception of modelling SBB use as a time-varying covariate, the approaches did not account for the possibility that individuals may have stopped treatment. We also assumed that treatment prescriptions directly represent treatment use, which may have

Table 2: Baseline hazards and regression coefficients of all models. Standards errors are presented alongside regression coefficients.

Model 1: treatment ignored; Model 2: SBB users excluded from the cohort; Model 3: SBB users censored at the date of treatment; Model 4: IPTW applied after censoring of treated patients; Model 5: SBB drop-in included as a binary covariate in the model; Model 6: SBB drop-in included as a time-varying covariate in the model; Model 7: marginal structural model.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
5-year baseline hazard	0.16 (0.01)	0.18 (0.01)	0.15 (0.01)	0.13 (0.01)	0.16 (0.01)	0.15 (0.01)	0.13 (0.01)
Age (years)	0.08 (0.00)	0.08 (0.01)	0.08 (0.01)	0.08 (0.01)	0.08 (0.00)	0.08 (0.00)	0.08 (0.01)
Female sex*	-0.43 (0.09)	-0.42 (0.10)	-0.47 (0.10)	-0.44 (0.10)	-0.41 (0.09)	-0.43 (0.09)	-0.44 (0.10)
Former smoker*	-0.63 (0.09)	-0.56 (0.10)	-0.57 (0.10)	-0.58 (0.10)	-0.61 (0.09)	-0.64 (0.09)	-0.63 (0.11)
Never smoker*	-0.81 (0.16)	-0.72 (0.17)	-0.72 (0.17)	-0.67 (0.17)	-0.80 (0.16)	-0.81 (0.16)	-0.71 (0.18)
Hypertension	-0.45 (0.10)	-0.43 (0.11)	-0.52 (0.11)	-0.48 (0.10)	-0.38 (0.10)	-0.47 (0.10)	-0.47 (0.10)
Diabetes	-0.04 (0.11)	0.04 (0.12)	-0.08 (0.13)	-0.02 (0.13)	0.02 (0.11)	-0.06 (0.11)	0.00 (0.13)
Angina pectoris	-0.22 (0.12)	-0.08 (0.14)	-0.27 (0.14)	-0.21 (0.13)	-0.09 (0.12)	-0.26 (0.12)	-0.22 (0.14)
Myocardial infarction	0.28 (0.17)	0.33 (0.19)	0.24 (0.19)	0.24 (0.22)	0.34 (0.17)	0.28 (0.17)	0.29 (0.20)
Atrial fibrillation	0.07 (0.12)	0.06 (0.13)	0.02 (0.13)	0.04 (0.13)	0.09 (0.12)	0.08 (0.12)	0.12 (0.13)
Heart failure	0.61 (0.09)	0.66 (0.10)	0.62 (0.10)	0.64 (0.11)	0.63 (0.09)	0.61 (0.09)	0.55 (0.10)
Stroke	0.19 (0.13)	0.16 (0.15)	0.10 (0.15)	0.07 (0.14)	0.22 (0.13)	0.17 (0.13)	0.10 (0.14)
Peripheral arterial disease	0.17 (0.15)	0.13 (0.17)	0.00 (0.17)	0.08 (0.18)	0.20 (0.15)	0.18 (0.15)	0.19 (0.17)
Baseline B01	0.02 (0.14)	0.07 (0.16)	0.11 (0.15)	0.11 (0.16)	0.03 (0.14)	0.02 (0.14)	-0.12 (0.19)
Baseline C03	0.32 (0.12)	0.36 (0.13)	0.37 (0.13)	0.36 (0.14)	0.33 (0.12)	0.32 (0.12)	0.23 (0.15)
Baseline C09	0.20 (0.16)	0.10 (0.18)	0.10 (0.18)	0.04 (0.18)	0.18 (0.16)	0.21 (0.16)	0.23 (0.18)
Baseline C10	0.18 (0.24)	0.01 (0.27)	0.15 (0.27)	0.15 (0.27)	0.11 (0.24)	0.19 (0.24)	0.22 (0.25)
Baseline H03	0.76 (0.16)	0.92 (0.16)	0.78 (0.17)	0.82 (0.20)	0.80 (0.16)	0.76 (0.16)	0.81 (0.21)
Baseline R03	-0.06 (0.11)	0.03 (0.11)	0.02 (0.12)	0.01 (0.12)	-0.05 (0.11)	-0.06 (0.11)	-0.06 (0.13)

Figure 2: 5-year mortality risk predictions provided by the seven models for patients in the development data set.

Predictions using models 2-7 are plotted against predictions provided by model 1 (treatment ignored). Blue points represent predictions for SBB non-users in the cohort; red points represent predictions for SBB users.

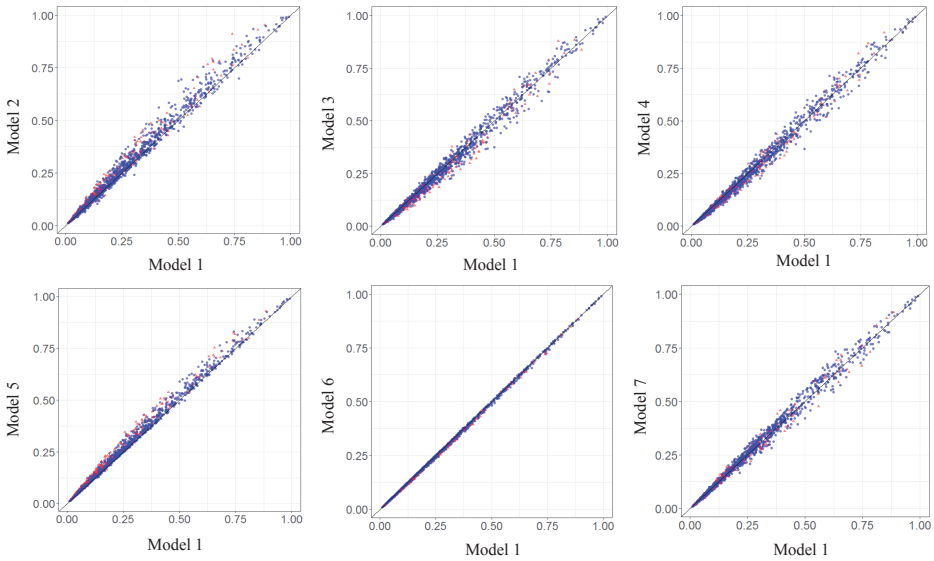
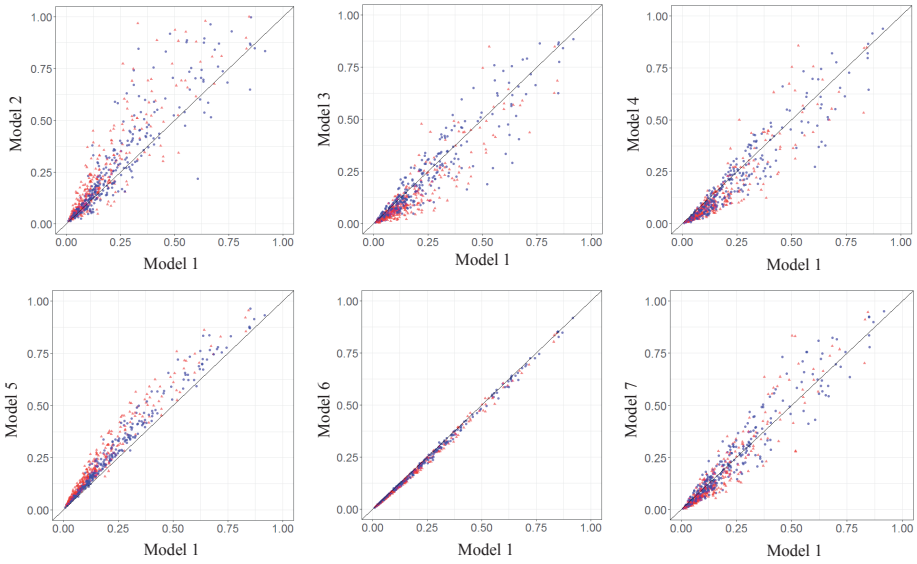


Figure 3: 5-year mortality risk predictions provided by the seven models for patients in a "highly treated" subset of the development data (50% treatment drop-in).

Predictions using models 2-7 are plotted against predictions provided by model 1 (treatment ignored). Blue points represent predictions for SBB non-users in the cohort ; red points represent predictions for SBB users.



overestimated the total amount of actual treatment use. Indeed, when the analysis set was sampled such that the SBB drop-in rate was 50%, the effects of removing or modelling treatment on risk predictions provided by the models became more apparent.

The limited effects of the more advanced IPTW and time-varying covariate approaches may also be due to residual confounding. While this study has demonstrated that it is feasible to account for the time-varying use of a pharmacological treatment in practice using prescription information, due to a lack of available information on disease severity and comorbidity during follow-up, we were not able to incorporate this important factor in our analyses. Unobserved confounding of predictor-treatment associations can bias IPTW estimates, as discussed elsewhere,[17] and as indicated by Figure 1, may result in a failure to remove the association between predictors and treatment. In such a case, these methods would not correctly estimates $Pr(Y^{T=0} = 1 \mid P)$. This highlights the challenge of implementing more complex approaches when using observational data with limited data collection.

Our structured approach to considering the potential bias caused by treatments falls in line with the ideas recently described by Sperrin et al.,[4] who recommended using MSMs to address treatment use for prognostic model development. This case study additionally considers approaches to exclude treated individuals or to directly model treatment as a time-varying variable in the prognostic model, both providing potentially less complex solutions than a MSM approach. Although all of the approaches considered in this paper attempt to remove the influence of treatment on predictor-outcome associations in a prognostic model, subtle differences between them may affect their suitability. First, as previously discussed,[4,22] an advantage of the MSM approach is that, unlike simply modelling treatment as a time-varying covariate, it can account for the association between prior treatment use and time-dependent confounders. The extent to which this translates to improved prognostic model performance in practice requires further investigation. In addition, both modelling treatment as a time-varying covariate, and the MSM approach may be preferred over censoring individuals and using IPTW, as these approaches utilize all available data and can readily account for treatment discontinuation after initiation.

This study has limitations. Ideally, to compare the impact of following the different modelling approaches, the resulting models should be compared in a truly external validation data set. As is often the case when trying to validate a prognostic model,[3] a validation data set without treatment use is not available, and thus we evaluated model performance in an untreated subset of the development cohort. Future studies using real data may need to consider employing advanced methods to account for treatment use in the validation data set.[21] Alternatively, future research could evaluate model

performance in a dataset with outcomes simulated to be in the absence of treatment. In addition, in this study we selected the parameters in our prognostic model based on clinical knowledge and a backwards selection procedure. It is not yet established whether shrinkage methods for model selection could influence the effect of treatment on a prediction model or vice versa, and requires further investigation. Finally, we present a single case study, and therefore the results should not be used to determine whether one modelling approach is always superior to another. This case study demonstrates that the choice of approach certainly affects the risk predictions made by a developed prognostic model. Whether these approaches are necessary for a given study will depend on factors such as the strength of the treatment effect, the duration of treatment, the number of individuals treated, and the strength of any associations between predictors in the model and treatment use – all of which need to be investigated in additional case studies and simulation studies.

In conclusion, this methodological case study has shown that different methods to account for time-varying treatment use when developing a prognostic prediction model using observational data result in different model coefficients and risk predictions. Limited differences were observed in terms of model discrimination or calibration. The effects of (time-varying) treatment use should nonetheless always be considered when designing data collection and analyses in a prognostic model development study. Only then can it be determined whether additional methods will be necessary to make adjustments for any treatment effects during model development and validation.

References

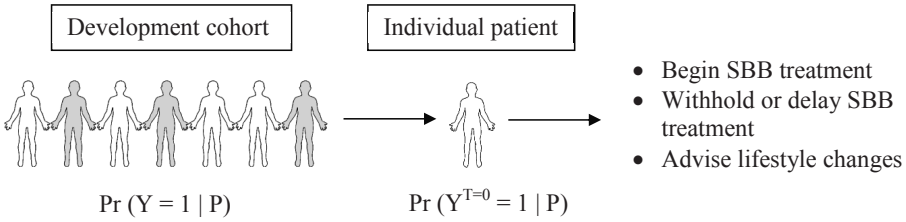
1. Liew SM, Doust J, Glasziou P. Cardiovascular risk scores do not account for the effect of treatment: a review. Heart (British Cardiac Society). 2011;97(9):689-97.
2. Groenwold RH, Moons KG, Pajouheshnia R, Altman DG, Collins GS, Debray TP, et al. Explicit inclusion of treatment in prognostic modelling was recommended in observational and randomised settings. Journal of clinical epidemiology. 2016.
3. 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.
4. Sperrin M, Martin G, Van Staa T, Peek N, Buchan I. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. arXiv preprint arXiv:170906859. 2017.
5. Liew SM, Doust J, Glasziou P. Systematic review did not consider problem of treatment effects. BMJ : British Medical Journal. 2012;345.
6. Cheong-See F, Allotey J, Marlin N, Mol BW, Schuit E, Ter Riet G, et al. Prediction models in obstetrics: understanding the treatment paradox and potential solutions to the threat it poses. BJOG : an international journal of obstetrics and gynaecology. 2016;123(7):1060-4.
7. 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. Annals of internal medicine. 2015;162(1):W1-W73.
8. Etminan M, Jafari S, Carleton B, FitzGerald JM. Beta-blocker use and COPD mortality: a systematic review and meta-analysis. BMC Pulmonary Medicine. 2012;12:48-.
9. Du Q, Sun Y, Ding N, Lu L, Chen Y. Beta-Blockers Reduced the Risk of Mortality and Exacerbation in Patients with COPD: A Meta-Analysis of Observational Studies. PloS one. 2014;9(11):e113048.
10. Rutten FH, Moons KG, Cramer MJ, Grobbee DE, Zuithoff NP, Lammers JW, et al. Recognising heart failure in elderly patients with stable chronic obstructive pulmonary disease in primary care: cross sectional diagnostic study. BMJ (Clinical research ed). 2005;331(7529):1379.
11. Rutten FH, Zuithoff NP, Hak E, Grobbee DE, Hoes AW. Beta-blockers may reduce mortality and risk of exacerbations in patients with chronic obstructive pulmonary disease. Archives of internal medicine. 2010;170(10):880-7.
12. Buuren Sv, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2010:1-68.
13. Esteban C, Quintana JM, Aburto M, Moraza J, Egurrola M, España PP, et al. Predictors of Mortality in Patients with Stable COPD. Journal of General Internal Medicine. 2008;23(11):1829-34.
14. Singanayagam A, Schembri S, Chalmers JD. Predictors of Mortality in Hospitalized Adults with Acute Exacerbation of Chronic Obstructive Pulmonary Disease. A Systematic Review and Meta-analysis. Annals of the American Thoracic Society. 2013;10(2):81-9.
15. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology (Cambridge, Mass). 2004;15(5):615-25.

16. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in medicine. 2015;34(28):3661-79.
17. Cole SR, Hernán MA. Constructing Inverse Probability Weights for Marginal Structural Models. American journal of epidemiology. 2008;168(6):656-64.
18. Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, et al. Adjusting for treatment switching in randomised controlled trials - A simulation study and a simplified two-stage method. Statistical methods in medical research. 2017;26(2):724-51.
19. Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ (Clinical research ed). 2010;340:b5087.
20. Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. Annual review of public health. 1999;20(1):145-57.
21. Pajouheshnia R, Peelen LM, Moons KGM, Reitsma JB, Groenwold RHH. Accounting for treatment use when validating a prognostic model: a simulation study. BMC medical research methodology. 2017;17:103.
22. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology (Cambridge, Mass). 2000;11(5):550-60.

Supplemental material

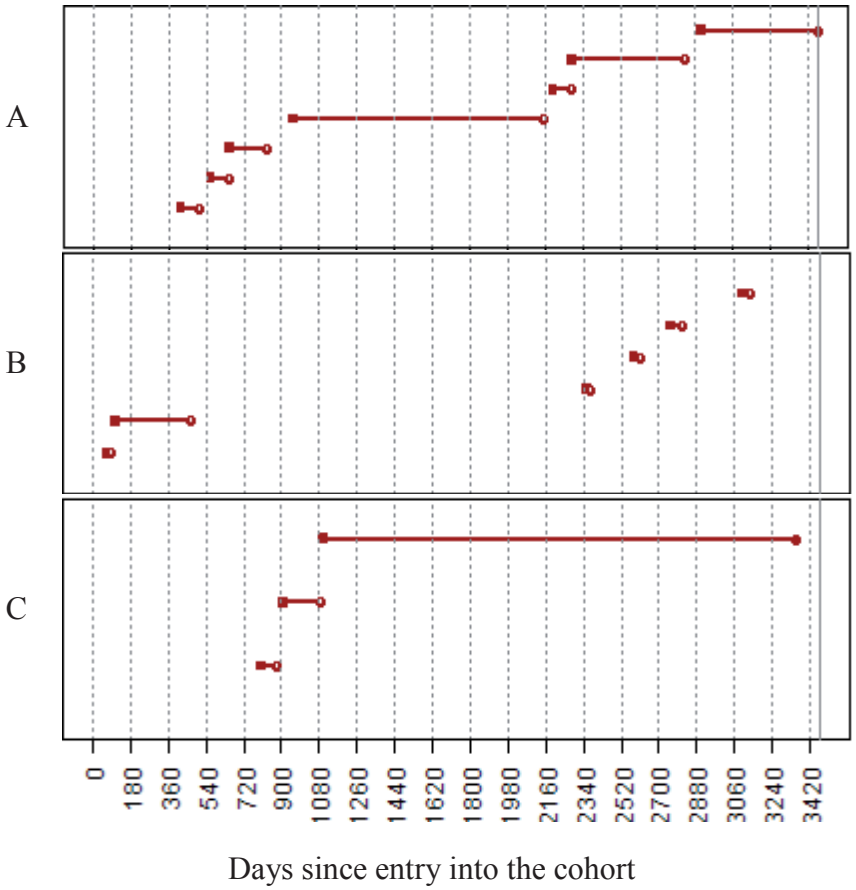
Supplemental figure 1: A summary of the case study objective.

A model developed using data on treated (grey) and untreated (white) individuals does not predict $Pr(Y^{T=0} = 1 \mid P)$, if treatment is not taken into account, and thus would not be suitable for guiding treatment decisions. In this case study, $Y^{T=0}$ represents the counterfactual outcome 5-year mortality if patients remain untreated with cardio-selective β-blocking agents (SBBs).



Supplemental figure 2: Selective β-blocker prescription patterns of three individuals.

Red lines represent separate periods where the patients had been prescribed SBBs. Patients A (seven periods of treatment) and C (three periods of treatment) were prescribed SBBs for the majority of follow-up. Patient B (six periods of treatment) had an irregular prescription profile, with a gap of nearly 5 years between two treatment periods.



CHAPTER 6

How variation in predictor measurement affects the discriminative ability and transportability of a prediction model

Accepted for publication
Pajouheshnia R, Van Smeden M, Peelen LM, Groenwold RH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. Journal of Clinical Epidemiology. 2018.

Abstract

Background

Diagnostic and prognostic prediction models often perform poorly when externally validated. We investigate how differences in the measurement of predictors across settings affect the discriminative power and transportability of a prediction model.

Methods

Differences in predictor measurement between data sets can be described formally using a measurement error taxonomy. Using this taxonomy, we derive an expression relating variation in the measurement of a continuous predictor to the area under the ROC curve (AUC) of a logistic regression prediction model. This expression is used to demonstrate how variation in measurements across settings affects the out-of-sample discriminative ability of a prediction model. We illustrate these findings with a diagnostic prediction model using example data of patients suspected of having deep venous thrombosis.

Results

When a predictor, such as D-dimer, is measured with more noise in one setting compared to another, which we conceptualize as a difference in "classical" measurement error, the expected value of the AUC decreases. In contrast, constant, "structural" measurement error does not affect the AUC of a logistic regression model provided the magnitude of the error is the same among cases and non-cases. As the differences in measurement methods between settings (and in turn differences in measurement error structures) become more complex, it becomes increasingly difficult to predict how the AUC will differ between settings.

Conclusion

When a prediction model is applied to a different setting to the one in which it was developed, its discriminative ability can decrease or even increase if the magnitude or structure of the errors in predictor measurements differ between the two settings. This provides an important starting point for researchers to better understand how differences in measurement methods can affect the performance of a prediction model when externally validating or implementing it in practice.

Introduction

Before prediction models are implemented in clinical practice, they should be externally validated, i.e. tested in individuals who were not a part of the data set used to develop the model.[1-4] Ideally, a model should perform well in terms of its discriminative ability and calibration [5] when validated in new sets of patients from different settings, e.g. from different clinical settings, geographical locations, or time periods. However, prediction models commonly perform differently - generally poorer - in new settings compared to what was observed in the development data set.[6] We then say that the transportability of the prediction model is low. Notably, failure for a model to transport well across settings indicates that the model cannot be readily implemented for new individuals.[7] Therefore, it is important that we understand what causes a prediction model to perform differently across settings. Discussions about variation in performance across data sets often focus on differences in patient characteristics.[8-10] Herein, we argue that variation in prediction model performance can also be explained (in part) by differences in how predictors are measured across settings, regardless of whether patient characteristics are similar or different.

The way that predictors are measured often varies from the development setting to validation or implementation settings. This occurs when predictor values are determined using different methodologies, protocols (e.g. fasting vs. non-fasting cholesterol measurements) or equipment, are measured by different people with varying levels of training, or are directly measured in one setting and measured by patient recall in a different setting, for example. Surrogate values for predictors may also be used when measurements of a certain predictor are unavailable in a data set.[11] Altogether, this can have a large impact on the value of measurements for individual patients; the value of a blood pressure reading, for example, is known to vary greatly depending on how the measurement is taken.[12] Therefore, we can expect that differences in the distribution of predictor values across different studies are not only due to true variation in the characteristics of patients but also the ways that their characteristics were measured.

In this report we describe how the discriminative ability of a prediction model varies across settings with variation in the measurement of predictors, and illustrate the effect both numerically and in a case-study about a diagnostic prediction model for deep venous thrombosis.

How the AUC is related to measurements of a predictor

Relating the AUC to the distribution of a continuous predictor

The area under the receiver operating characteristic curve (AUC) indicates how well a prediction model can discriminate between individuals who have/will have (cases) or do not have/will not have (non-cases) the health outcome of interest (e.g. disease or health state).[13] Assuming a continuous predictor (which may be a linear combination of several predictors) follows a normal distribution among cases and non-cases separately, it has been shown that the AUC of a predictor of a binary outcome can be approximated by [14]:

$$\text{AUC}=\Phi\left(\frac{\mu_1-\mu_0}{\sqrt{\sigma_1^2+\sigma_0^2}}\right) \qquad [1]$$

Here μ_1, μ_0, σ_1^2 and σ_0^2 refer to the means and variances of the predictor in the cases (μ_1, σ_1^2) and non-cases (μ_0,σ_0^2), and Φ denotes the cumulative normal distribution function. From [1] we can see that the AUC is a function of the mean and variance of the values of a predictor that are observed for cases and non-cases.

Relating the AUC to the distribution of a predictor measured with error

To understand how the measurement of a predictor can affect the AUC of a prediction model, we turn to an existing taxonomy for measurement error (for further details see [15-17]). First consider a candidate predictor, for example height. In one sample, the height of patients is measured directly by a research assistant, providing an accurate measure of heights in the sample. In another sample, height is self-reported. Given that patients are likely to recall their height with a certain amount of error, the self-reported height value observed for an individual i (W_i) represents their accurately measured height (X_i) plus some additional error (U_i) [15]:

$$W_i = X_i + U_i. \qquad [2]$$

A common model for U_i is the "classical" measurement error model where U follows a normal distribution with a mean value of zero, and a (constant) variance, τ. Under this model, the error in self-reported height is considered random and on average the measurements are unbiased ($\text{E}(W) = \text{E}(X)$) but have additional variance, such that the expected variance of W is equal to the sum of σ^2 (the variance of the accurately measured predictor X) and τ. It follows that the expected value of the AUC of the predictor in the sample, measured with random error, is:

$$\text{AUC}=\Phi\left(\frac{\mu_1-\mu_0}{\sqrt{\sigma_1^2+\sigma_0^2+2\tau}}\right). \qquad [3]$$

From this, we can see that as the amount of random error with which a predictor is measured increases, the discriminative ability of the predictor is expected to decrease, provided the other parameters in [3] remain constant.

A general expression relating measurement error to the AUC of a continuous predictor

Measurement error can also affect the mean of the observed values, which may also vary between cases and non-cases. Expression [2] can be extended as such:

$$W_i = \Psi_y + X_i\theta_y + \varepsilon_i, \qquad [4]$$

where y is an indicator to distinguish between cases (y=1) and non-cases (y=0). Further, we assume $\varepsilon|y \sim N(0,\tau_y)$. The mean and variance components of the observed predictor values in the cases and non-cases can now be defined. Let the expected values of the predictor X be defined as $\overline{X}_0=E(X|Y=0)$, $\overline{X}_1=E(X|Y=1)$, and similarly, the values for error-contaminated predictor W, be defined as $\overline{W}_0=E(W|Y=0)$, $\overline{W}_1=E(W|Y=1)$. Hence, in expectation,

$$\overline{W}_1=\Psi_1 + \theta_1 X_1 \qquad [5]$$
$$\overline{W}_0=\Psi_0 + \theta_0 X_0 \qquad [6]$$
$$\Sigma_1^2=\sigma_1^2\theta_1^2+ \tau_1 \qquad [7]$$
$$\Sigma_0^2=\sigma_0^2\theta_0^2+ \tau_0 \qquad [8]$$

It follows that, under the same conditions required for expression [1] to hold, this expression can be extended to incorporate measurement error by substituting the means and variances of predictor X in the cases and non-cases for values of predictor W (measured with error), such that

$$\text{AUC}=\Phi\left(\frac{\overline{W}_1-\overline{W}_0}{\sqrt{\Sigma_1^2+\Sigma_0^2}}\right) \qquad [9]$$

Differences in measurement error between settings lead to changes in the AUC

As explained, the way that predictors are measured in samples from different settings varies, which could result in variation in measurement error. Notably, it follows from expression [9] that if differences in measurement between settings translate to differences in the structure or magnitude of the measurement error associated with the predictors, the AUC can vary across these settings. Figure 1 (scenarios S1 and S2) shows that when the amount of random error across settings increases, the expected value of the AUC decreases. In contrast, depending on the direction of the error, and whether it is present equally in the measurements of both cases and non-cases, non-random error can cause the AUC to increase, decrease or may have little effect. Therefore we see that for the discriminative ability of a predictor to transport to a new setting, the measurement error of that predictor must also be transportable. Furthermore, the mean and variance of the predictors in the absence of measurement error remained constant across scenarios. In reality it becomes extremely challenging to predict how the AUC will change across settings, because the AUC is a function of predictor means, variances and error, all of which can change between settings.

Case-study: differences in measurements from development to validation

Data from 1295 patients with possible deep venous thrombosis (DVT) [18] were used to examine how differences in the measurement of a predictor across samples affect the discriminative ability of a diagnostic prediction model. First, a model to predict the presence of DVT was developed with a single predictor, D-dimer (log-transformed, continuous measurements of the biomarker were used), using logistic regression on a random half of the data (a split-sample procedure for illustration purposes). Next, we explored how measuring D-dimer with greater error in a validation sample could affect its discriminative power. Error of increasing magnitude was simulated and added to the D-dimer measurements in the remaining half of the data, and subsequently the AUC was calculated in this half of the data. The AUCs reported in Figure 2 denote the average AUC after replicating the entire procedure (from data splitting to calculating the AUC) 1000 times. Figure 2 shows that when measurements were conducted less accurately (i.e., with increasing amounts of noise or "classical error", see section 3) there was greater overlap between the distributions of the predictor values of the cases and non-cases in the validation sample. This translated to a strong reduction of the AUC, from 0.89 in the development sample to 0.67 in the validation sample with a 200% increase in log D-dimer variance relative to the actual variance. In contrast, a fixed increase in the D-dimer measurements (constant "structural error", see section 3) caused a uniform shift in the predictor values and thus the AUC remained unchanged with increasing

error, as was also seen in Figure 1, scenario 3. Finally, the prediction model including D-dimer was extended with additional predictors: sex, oral contraceptive use, presence of a malignancy, recent surgery, absence of leg trauma, vein distension and the difference between calf circumferences, to reflect a published diagnostic model. [18] All predictors except D-dimer were assumed to be measured in the same fashion in the development and validation samples. The same trends were observed as in the univariable (D-dimer only) model; the AUC ranged from 0.90 in the development set to 0.70 in the validation sample with a 200% increase in log D-dimer variance relative to the actual variance, and remained stable with fixed increases in the D-dimer measurements.

Figure 1: The effect of measurement error in a single continuous predictor on the AUC when predicting a binary outcome.

Data of size n=10^5 were simulated such that the values of a continuous predictor X, measured without error, were normally distributed for non-cases ($X_0 \sim N(1,0.5^2)$) and cases ($X_1 \sim N(2,0.5^2)$). Measurement error was simulated on top of the error-free measurements according to expression [4], such that values, W, were observed instead of the true values of X, and the AUC was estimated using expression [9]. Each point on the curves represents a sample with a certain amount and type of measurement error. The horizontal axes represent the measurement error parameter values used to vary the observed values, W, using expressions [5-8]. In scenario S1, random (classical) error was added by increasing the random error term τ. In scenario S2, random error was only added to the cases (τ_1). In scenarios S3-S5, a constant (structural) error value, Ψ, was added to the measurements- to both cases and non-cases (S3), cases only (Ψ_1) (S4), and non-cases only (Ψ_0) (S5), respectively. In scenarios S6-S8, error proportional to the true value of X, θ, was added - to the cases and non-cases (S6), cases only (θ_1) (S7), and non-cases only (θ_0) (S8), respectively.

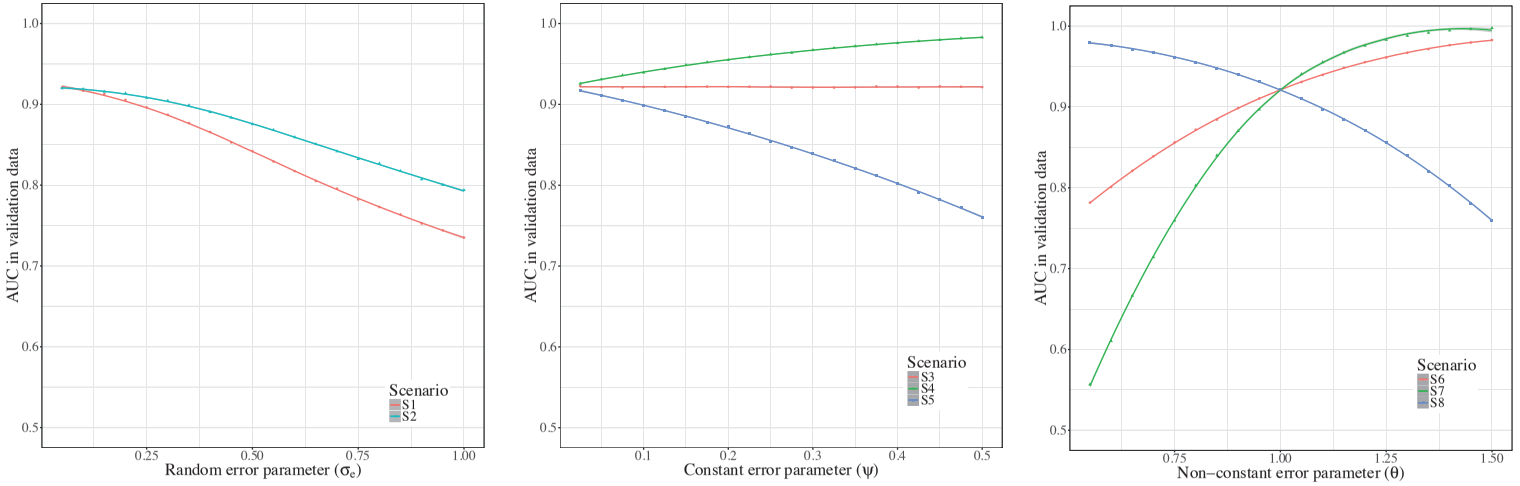
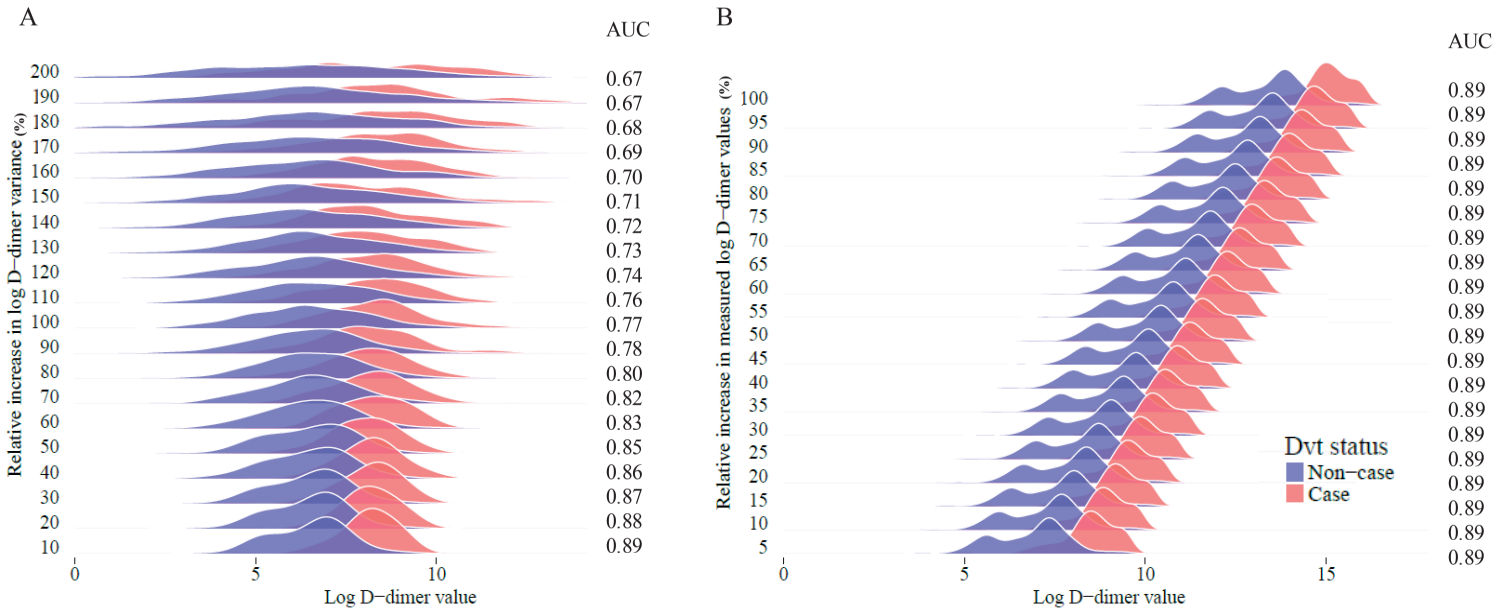


Figure 2A-B: The transportability of the discriminative value of D-dimer for separating DVT cases from non-cases, from the development sample to validation samples with increasing amounts of measurement error.

The DVT data were randomly split 50:50, simulated error was added to the DVT values measured in the validation sample and the AUC was calculated. The process was repeated 1000 times and the average of the AUCs was calculated. A: random (classical) error, where the error was randomly sampled from a normal distribution ($\varepsilon \sim N(0, \sigma_{D\text{-}dimer} * m)$, and m ranged from 0.1-2. B: constant (structural) error, such that $\varepsilon = \widehat{X}_{D\text{-}dimer} + m$, and m ranged from 0.05-1.



Concluding remarks

Differences in the way predictors are measured across settings can cause the discriminative ability of a prediction model to appear to be worse, but perhaps surprisingly can appear to be better as well, in one setting compared to another. Thus, for a prediction model to transport well to new patient samples, i.e. from development, to validation and finally implementation in daily practice, measurements methods should be comparable across each sample. We propose that differences in the way a predictor is measured across samples from different settings can be viewed in the context of measurement error. Prediction does not require the "true" value of a variable to be measured, rather, predictions are made using observed measurements.[15] Whether predictor measurements deviate from their "true" (e.g. biological) value becomes important only if this deviation from "truth" varies from one sample to another.

A number of studies have investigated factors that influence the performance of a prediction model. The effect of measurement error on prediction model performance within a single sample has been examined elsewhere.[19-20] Others have investigated how correlation between predictors in a model is related to model performance. A simulation study by Kundu et al. [21] found that differences in the correlations between predictors in validation samples compared to the development sample can result in differences in the AUC. Given that differences in measurement methods can affect the variance of predictor measurement in a sample, and the correlation between two variables is a function of their variances and covariance, our findings explain how differences in correlations can arise between samples, and how this can affect the AUC of a model.

Variation in model performance due to differences in how predictors have been measured can have different implications. First, differences in performance can arise when predictors in a validation sample have been measured using methods that do not reflect current standards. This could happen if the validation sample comes from an historical cohort, in which outdated methods of measurement were used. Alternatively, if data were collected in a highly protocol-driven setting, such as for a randomized trial, measurements may be more precise than in clinical practice. In such cases, evidence of poor model transportability is weakened by non-representative measurements in the validation sample. Second, differences in measurements from development to validation might reflect true variation in clinical practice. In this case, we might conclude that poor performance in a validation sample is evidence of limited transportability of the model. Finally, it could be that the model itself is outdated, and since its development, the methods used to measure a predictor have improved. Poor performance in a contemporary validation sample could indicate that the model requires updating.

We have presented a starting point for the exploration of the impact of differences in measurements on prediction model performance, but further attention is required in a number of areas. First, the mathematical relationship we present between measurement error and model discrimination is restricted to the case of a single continuous predictor, and is based on strict assumptions. Future research could use computer simulations to further explore the impact of differences in measurements on the discriminative ability of multivariable prediction models across samples, and could investigate the misclassification of categorical predictors. Second, while we have discussed model discrimination, model calibration requires separate attention. Finally, we do not comment on the use of correction methods for measurement error when developing or validating a prediction model, as this remains a topic for further investigation.

To conclude, if the measurement of predictors varies from sample to sample, we can anticipate changes in the discriminative ability of the model. Discussions about variation in prediction model performance across settings should therefore also consider variation in predictors measurements. The way predictors are measured when developing or validating a prediction model should mimic the way predictors are measured in practice in order to obtain realistic and relevant estimates of prediction model performance.

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. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19(4):453-73.
4. McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. Jama. 2000;284(1):79-84.
5. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating: Springer Science & Business Media; 2008.
6. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. Journal of clinical epidemiology. 2015;68(1):25-34.
7. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Annals of internal medicine. 1999;130(6):515-24.
8. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. American journal of epidemiology. 2010;172(8):971-80.
9. 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. Journal of clinical epidemiology. 2015;68(3):279-89.
10. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. Bmj. 2016;353:i3139.
11. Smith T, Muller DC, Moons KGM, Cross AJ, Johansson M, Ferrari P, et al. Comparison of prognostic models to predict the occurrence of colorectal cancer in asymptomatic individuals: a systematic literature review and external validation in the EPIC and UK Biobank prospective cohort studies. Gut. 2018.
12. Handler J. The importance of accurate blood pressure measurement. The Permanente Journal. 2009;13(3):51.
13. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36.
14. 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.
15. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models: a modern perspective. 2 ed: Chapman \& Hall /CRC Press; 2006.

16. Gustafson P. Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments: Chapman and Hall/CRC; 2004.
17. Buonaccorsi J. Measurement Error: Models, Methods and Applications. 2010.
18. Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost. 2005;94(1):200-5.
19. Khudyakov P, Gorfine M, Zucker D, Spiegelman D. The impact of covariate measurement error on risk prediction. Statistics in medicine. 2015;34(15):2353-67.
20. Rosella LC, Corey P, Stukel TA, Mustard C, Hux J, Manuel DG. The influence of measurement error on calibration, discrimination, and overall estimation of a risk prediction model. Population health metrics. 2012;10(1):20.
21. Kundu S, Mazumdar M, Ferket B. Impact of correlation of predictors on discrimination of risk models in development and external populations. BMC Medical Research Methodology. 2017;17:63.

CHAPTER 7

Reducing research waste: When and how to use data from randomized trials to develop or validate prognostic prediction models

Abstract

Prediction models have become an integral part of clinical practice, providing information for patients and clinicians and support for clinical decision making. The development and validation of prognostic prediction models requires substantial volumes of high-quality information on relevant predictors and patient health outcomes. Primary data collection for prognostic model research comes with substantial costs and if suitable data are already available for prognostic modelling, performing such a new study can be seen as a form of research waste. Randomized clinical trials (RCTs) are a source of high-quality clinical data with a largely untapped potential for use in further research. This article addresses when and how data from a RCT can be used additionally for prognostic model research and guidance is provided to help researchers with access to RCT data to evaluate the suitability of their data for the development and validation of prognostic prediction models.

Summary points

- To minimize research waste, data from randomized clinical trials (RCTs) might be considered for the development or validation of a prognostic prediction model.
- Advantages of RCT data include completeness, accuracy and consistency of the data, and detailed protocols.
- Randomized treatment allocation facilitates the development and validation of prognostic models that predict risk in the presence or absence of (a certain) treatment.
- RCT data might be less suitable due to selective patient or centre inclusion, extraneous trial effects or overly specialized predictor measurement, which all may limit the generalizability of prognostic models to real-life practice. Other limitations might be too short-term or clinically irrelevant, surrogate outcomes, or an insufficient sample size for prognostic model development or validation.
- This paper provides guidance to appraise the suitability of RCT data for prediction research by examining both potential benefits and limitations.

Introduction

Prediction model research requires substantial amounts of high-quality clinical data. Although prospective data collection designed specifically to develop or validate a prognostic prediction model is typically advocated,[1] this is often not feasible or desirable due to the vast costs involved. Randomized clinical trials (RCTs) provide a tempting alternative data source for the development and validation of prognostic prediction models: in the last twelve months alone, nearly 20,000 RCTs of therapeutic interventions were published, generating an unfathomable amount of data (see Supplement for search query). Yet the valuable information gathered in RCTs remains largely untapped by the research community and one could view this as a tragic source of research waste. At the same time, despite the widespread belief that RCTs are the "gold standard" for data generation, their suitability for addressing questions of a descriptive (i.e. predictive) nature has been questioned.[2] This article starts from the perspective that we would like to develop or externally validate a prognostic prediction model and we have access to individual participant data- herein referred to as "data"- from a relevant phase III (or possibly IIb) RCT. We present the opportunities that RCT data can offer, describe potential limitations that must be considered, and navigate the "dos and don'ts" of developing or externally validating a prognostic prediction model with RCT data.

Opportunities when using RCT data

To date, a number of prognostic prediction models have been effectively developed and validated using RCT data (see Table 1). Data generated by a RCT may confer certain benefits over data from alternative sources, such as from predesigned, observational studies, electronic health records, disease-specific registers or administrative medical databases. We outline the key opportunities that RCT data might provide when developing or externally validating a prognostic prediction model.

O1. Data quality: completeness

The completeness of data from RCTs can be an important asset for prediction model studies. Missing data is a serious and almost ubiquitous issue for studies that develop or externally validate prediction models.[13,14] Although numerous methods exist to handle missing data, the best solution is undoubtedly its prevention. To develop a prediction model, one ideally needs complete information on all candidate predictor variables, measured in all individuals in the study.

Table 1: Ten examples of high-impact studies where a prognostic prediction model was developed or validated using data from a RCT

Prediction model	Clinical use	Data source	Study type	Sample size*	Ref.
IMPACT model	Risk of outcome after TBI	Eleven RCTs and three observational studies	Development	8509 (5748)	3
TIMI risk score	Risk of death or ischemic events in patients with UA/NSTEMI	Two RCTs (TIMI 11B, ESSENCE)	Development	1957 (327)	4
EORTC risk tables	Prediction of recurrence of stage Ta T1 bladder cancer	Seven RCTs (EORTC trials)	Development	2596 (1240)	5
ADVANCE cardio-vascular risk model	4-year cardiovascular disease risk in patients with type II diabetes	One RCT (ADVANCE)	Development	7168 (473)	6
S_2TOP-BLEED	Risk of major bleeding in patients with a TIA/ ischemic stroke on antiplatelet agents	Six RCTs (CAPRIE, ESPS-2, MATCH, CHARISMA, ESPRIT, PRoFESS)	Development	43112 (1530)	7
FRAX tool	10-year risk of fracture	One RCT	Validation	1422 (229)	8
MCL International Prognostic Index	Prediction of mantel cell leukaemia survival	Two RCTs (MCL Younger, MCL Elderly)	Validation	958 (316)	9
EFFECT model	Risk of mortality within one year of hospitalization for heart failure	One RCT (EVEREST)	Validation	2662 (712)	10
SYNTAX Score II model	Mortality prediction after PCI or CABG	Two RCTs (BEST, PRECOMBAT)	Validation	1480 (90)	11
OHTS-EGPS model	5-year risk of open-angle glaucoma	Two RCTs and two observational studies	Validation	1038 (105)	12

Abbreviations: TBI, traumatic brain injury; UA/NSTEMI, unstable angina/ non–ST-segment elevation myocardial infarction; PCI, percutaneous coronary intervention; CABG, coronary artery bypass graph. * Number of participants (number of events).

However, practically a model may be developed using whichever predictor variables are available. External validation requires complete information on all predictors of the model that is under evaluation.

Throughout the design and conduct of a RCT, a number of strategies may be employed to facilitate the complete collection of information on predictors and outcomes in all trial participants.[15] This may include the training of research staff before starting data collection and incentives for data collectors to record complete information. While this may be a challenge in multicentre trials, the existence of a common, shared protocol can help to maintain consistent data collection across the centres. A unique feature of RCTs is detailed study monitoring, usually by a number of separate committees.[16] Trial oversight committees, such as data monitoring committees, monitor the presence of missing data in a trial. These efforts work synergistically with central and on-site monitoring to keep track of missing data and prevent additional missing data.

In addition, RCT data can include detailed information on important post-baseline events, which could affect the prognosis of participants. Such details are often not available in observational databases. Post-baseline events such as changes in treatment, the use of rescue medications or competing outcome events may need to be accounted for when developing or externally validating a prediction model, and should be reported alongside the results.[17]

O2. Data quality: accuracy and consistency

Accurate and consistent predictor and outcome information is a requisite for accurate prediction models. RCTs are commonly regarded as a source of high-quality health data. As with data completeness, considerable amounts of time and money are channelled towards ensuring data are correctly measured and recorded.

First, adherence to the trial protocol and standard operating procedures facilitates the accurate and consistent measurement of predictors and outcomes, in particular for specific variables of interest in the trial (although this might not reflect actual variation in practice, see section L4). Second, case report forms require the recording of detailed clinical information and can help to prevent the recoding of impossible values, forming a part of the quality assurance process in a RCT.[18] Third, as with data completeness, study monitoring in RCTs helps to maintain accuracy and consistency in the recorded data. For example, central monitoring includes the checking of data for unusual patterns or implausible values.[16] In addition, source data verification and electronic data capture methods form an additional layer of data validation.[19] Finally, a centralized system for the adjudication of outcome events can be especially important when outcome

measurements are subjective. Altogether, these systems and processes can yield data that satisfy the quality requirements of prediction model studies.

O3. Protocol and records

The availability of a trial protocol permits a better understanding of how predictors and outcomes were defined and measured. This information is helpful in several respects. First, data on how and when predictors were measured may provide insight into the suitability of a predictor for inclusion in a prediction model. For example, if the protocol states that a certain predictor should be measured at a time point that is not relevant to routine clinical practice, one might not select the predictor. Second, knowledge of the operationalization of predictor or outcome measurements provides insight into how well a model may perform in practice, and can inform risk of bias assessment when reviewing a prediction model study.[20] In addition, information stored in such meta-data may also have predictive value. The timing of measurements, whether taken during the day or night, can be highly predictive of clinical outcomes.[21] Such information could also be used in statistical models to impute missing data.

O4. Treatments

Often in practice, clinicians are interested in the following question: "What will happen to my patient if I do or do not treat them?". By addressing this question, prediction models can be used to support clinical decisions, as well as provide information to clinicians and patients for counselling.[1] For this purpose, prediction models must predict risks for patients in the absence (or presence) of a certain treatment. This can prove challenging in non-RCT data due to the non-random use of treatments by patients,[22] and advanced statistical methods may be needed to correctly account for this.[23,24] In the case of RCT data, the effect of treatment use can be solved by simply developing or validating the prognostic model in the untreated (or placebo) trial arm or by including the randomized treatment as a predictor in the model, along with terms for any other treatment-predictor interactions (model development only).[22] However, it must be recognized that the placebo-arm of a placebo-controlled trial may not represent truly untreated patients in usual practice (see section below "L5. Extraneous trial effects").

Limitations and challenges when using RCT data

Available data from a RCT can suffer from a number of limitations which may reduce the viability of its use for prognostic prediction modelling. Below, we present key challenges when considering using RCT data for prognostic model research. Where necessary, the issues are addressed separately for model development and model validation.

L1. Consent

Consent to re-use RCT data for prediction modelling may not have been given by the trial participants. In contrast to data repositories established specifically for scientific research purposes (e.g. the UK Biobank[25]), which have very broad consent for data re-use,[26] trials may not always have asked for a sufficiently broad consent. However, compared with routinely collected data, which faces even greater challenges with consent in light of the recently implemented 2016 EU General Data Protection Regulation,[27] RCT data might prove more accessible, especially if trials begin to adopt broad consent for data re-use, as recommended.[28] It is likely that researchers will need to consult their institutional review board before using RCT data for secondary analysis, but whether this satisfies ethical and legal requirements needs further examination.

L2. Selective inclusion of centres

The centres that participate in RCTs might not be representative of clinical practice in general.[29] Specifically, generalizability of a prognostic prediction model could be limited if only specialist trial centres (e.g. academic medical centres) or experienced clinicians with high performance ratings were invited to participate.[30] In such cases, the associations between predictors and the outcome, and the incidence of outcomes could be different in the trial setting compared to routine clinical practice, of which both could affect the performance of a prognostic prediction model.[31]

L3. Selective eligibility and enrolment

RCTs commonly have narrow participant eligibility criteria, e.g. often excluding frail, multi-morbid or vulnerable patients.[32-36] At the same time, some of the most challenging clinical decisions are for these groups of patients. Thus, RCTs may not provide sufficient information for prediction research in these clinically relevant patient subgroups. When developing a prediction model using a selective subset of patients, we assume that the predictor effects and functional forms of their associations with an outcome are the same across the patient subsets included and excluded from the RCT. In addition, the participants invited to enter a RCT and those who actually enrol and remain in the trial until completion can substantially differ.[37] For example, the requirement of informed consent from participants has been shown to result differences between the patients enrolled and not enrolled in the a trial.[38,39] As with selective eligibility, this may limit the value of RCT data for prognostic model development. This may not be as problematic for external validation, but may limit the generalizability of results to broader patient populations.

L4. Predictor measurement

As mentioned, protocol-driven data collection by trained research staff can help to improve the accuracy and consistency of clinical measurements, which in turn can

improve the accuracy of a prediction model.[40] For the purpose of prediction, however, the measurement of predictors should closely reflect how they are measured in regular clinical practice. Thus the use of unrealistically accurate measurements – which may occur if specialist personal or equipment are used in a RCT - when developing a prediction model could reduce the transportability of the model to clinical practice, and the findings of a validation study may not represent how the model will truly perform in practice.[41]

L5. Extraneous trial effects

The effects of being enrolled in a trial on participant behaviour are well-documented.[42] Knowledge of enrolment in a trial can lead to participants behaving differently, even reporting more optimistic outcomes,[43,44] an effect commonly termed the "Hawthorne effect". The enrolment of a centre in a RCT may also affect the behaviour of healthcare professionals and as a result the prognosis of a patient enrolled in a trial may be better than it would have been according to routine care.[45] In the case of placebo-controlled trials, patients on placebo may also exhibit a placebo (or nocebo) effect, which may positively (or negatively) affect their outcomes.[46] In addition, the presence of a "protocol effect" or "care effect" - whereby adherence of centres to a strict trial protocol may improve patient outcomes (e.g. through additional monitoring) compared to patients not enrolled in the trial [47,48] - may hamper the generalizability of RCT data to clinical practice. Thus, a trial that suffers from strong extraneous effects might not provide suitable data for prediction research.

L6. Short-term and surrogate outcomes

Long-term, clinically relevant outcomes are often of interest in prognostication in daily practice. For example, models to predict cardiovascular disease risk are commonly designed to predict outcomes within 10-years.[49] Such models require very long follow-up, which is rarely available in large RCTs. However, unlike validating a model for long-term prediction, with short-term outcome data, which is not advisable, there may still be clinical use of a prediction model developed with short-term outcomes. In addition, RCTs often opt for "surrogate endpoints", to replace more costly long-term outcomes.[50] If a prediction model is to be used to inform patients and healthcare professionals, surrogate endpoints may have insufficient clinical relevance if the surrogate is imperfectly correlated with the clinical outcome, whether used to develop or validate a prediction model.

L7. Sample size

Prognostic model development research often requires substantial samples. While there is currently no consensus on the sample sizes required for prediction, evidence suggests that for the development of a logistic regression model, at least 20 outcome events [51]

(if not more [52]) are needed for every predictor included in the model. Similarly, reliable prediction model validation requires data samples with a minimum of several hundreds of outcome events[53]. Obviously, RCTs are not designed and powered with prediction research in mind. Thus data from a single RCT may simply not suffice. While approaches such as penalized regression can help to prevent the "overfitting" of prediction models in small data sets [54], large samples may yet be needed for modern modelling techniques [55]. Data from large, multicentre trials can, however, provide a solution to this issue. In addition, as seen in Table 1, the combination of individual participant data from more than one RCT can greatly increase the amount of available data.

How and when to use data from a RCT for prediction research

When data from a RCT are available, researchers must weigh the advantages (data quality, completeness, treated and untreated arms, and protocol) against any limitations, both described above. We suggest that the decision process can be separated as follows:

1. Criteria that must be met

- There must be acceptable patient consent (or under certain circumstances a waiver by an institutional review board [56]) for reuse of the RCT data for prediction research.

2. Criteria that may seriously limit the usefulness of the data

- Insufficient sample size or follow-up, or no availability of important predictors or outcomes will seriously limit the suitability of RCT data.

3. Criteria that may limit the usefulness of the data

- Selection of patients or centres, experimental effects and highly protocol-driven predictor measurements may all limit how representative the trial data are of the target population for the prediction model.

To aid in this process, Table 2 presents a series of questions to ask when assessing whether data from a certain RCT are suitable for developing or externally validating a prognostic prediction model. For each situation, general advice is provided to help researchers reach a decision. The decision to use a given set of data from a RCT will depend on the specific research question and remains largely subjective. In addition, to help gain an overall picture of the suitability of a RCT as a whole, researchers can benefit from constructing a diagram, such as in Figure 1. In this (fictional) example, consent for secondary use of the data was available and the data received a high "score" for this criterion. Following this, the remaining criteria were assessed. From this we see that the data might be a good candidate for an external validation study, but centre and participant selection may

limit the generalizability of a prognostic model developed using the data. With such a picture, researchers can decide whether the benefits of using the RCT data outweigh the limitations.

Table 2: How to assess whether data from a RCT is suitable for developing or externally validating a prediction model.

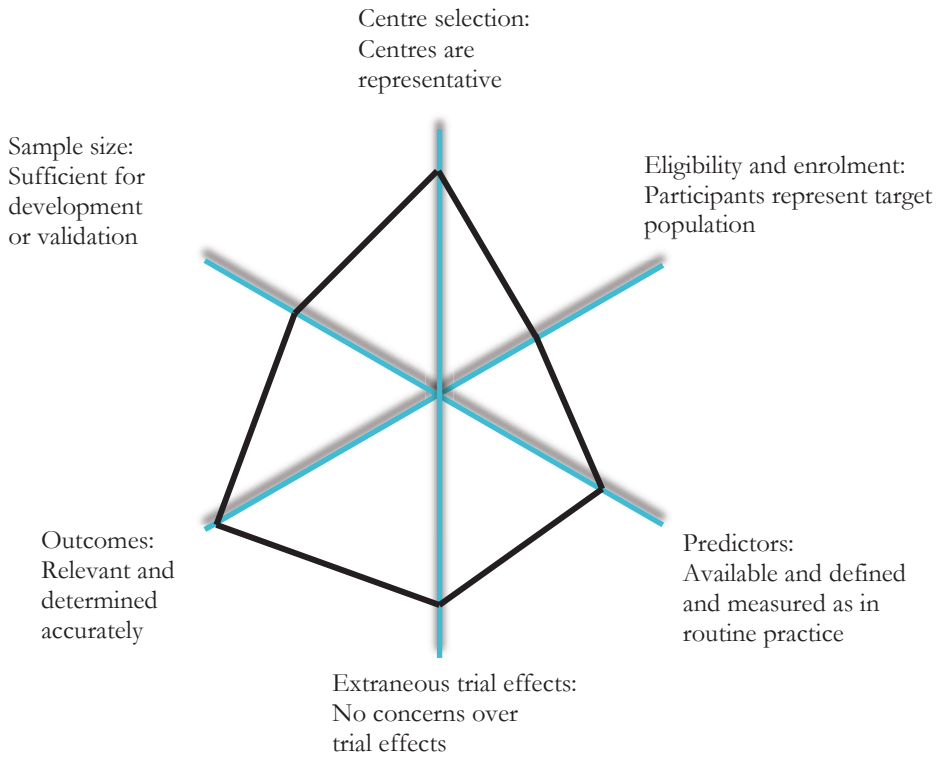
Suggestions are provided for how to proceed in the event that any of the considerations might seriously limit the usefulness of a RCT for prediction research.

Consideration	Questions to ask	How to proceed
1. Consent	Did patients give consent or has consent been waived for the data to be re-used for prediction research?	If consent for re-use is inadequate, data should not be used.
2. Selective inclusion of centres	Are centres (their expertise, facilities and use of complex interventions) in the trial representative of centres where you might use the prediction model?	*Development:* May not be suitable if trial centres do not represent standard practice. *External validation:* May still be used, but report limitations to the generalizability of results.
3. Selective eligibility and enrolment	Did the trial eligibility criteria result in the exclusion of relevant patient groups for the prediction model? Are patients who did/did not enrol (after invitation), and patients who remained/left the study comparable in terms of their characteristics and possible prognosis?	*Development:* May not be suitable if participants do not represent the target population for prediction *External validation:* May still be used, but report limitations to the generalizability of results.
4. Predictor measurement	Were predictors measured as they would be in routine clinical practice?	If the methods of predictor measurement are seriously unrealistic, data may not be suitable.
5. Extraneous trial effects	Could enrolment in the trial have influenced patient prognosis beyond a treatment effect?	If there is evidence for strong experimental effects, data may not be suitable.
6. Short-term and surrogate outcomes	Were the clinically relevant outcomes measured? Is there sufficiently long follow-up for outcomes?	If outcomes or the timing of their measurement are not relevant, data may not be suitable.
7. Sample size	How many participants were enrolled and remained in the study? What proportion had the outcome?	*Development*: consider methods for data with few events [54], but may be too small for any meaningful modelling. *External validation*: Data may not be suitable.

Finally, if a decision is made to use the RCT data, this information can be used when reporting any study limitations.

Figure 1: A graphical example of how the suitability of data from a RCT can be assessed after consent for use has been confirmed.

A researcher might attribute a level of confidence that the RCT does not suffer from the eight limitations described. For example, the black lines converge at the end of the "consent" axis, indicating that patient consent for re-use of the data for prediction was obtained. The greater the area within the black lines, the more suitable the data might be for prediction.



Concluding remarks

When data from a RCT are available for the secondary purpose of developing or validating a prediction model, the opportunities and limitations of these data require careful consideration. Available data from RCTs can, if used appropriately, be a viable substitute for costly and labour-intensive dedicated data collection for prediction research. In essence, by first recognizing the possibilities that RCT data offer and then

carefully appraising available data, we can maximize the chance of utilizing data that permit high-quality prognostic model research, while avoiding unnecessary, costly primary data collection.

Inevitably, there remain fundamental challenges that are universal to the secondary use of data for research, such as the systematic absence of data on certain key predictors. In circumstances such as these, researchers might consider designing a dedicated study to collect data to develop of externally validate a prediction model.

We hope that researchers will *cautiously* seize the opportunities that data generated by RCTs provide, to improve both the quality and efficiency of future prediction research.

Acknowledgements

We would like to thank Dr. Rieke van der Graaf for providing insight into ethics and consent for trial data reuse.

References

1. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338.
2. Vandenbroucke JP. Observational Research, Randomised Trials, and Two Views of Medical Science. PLOS Medicine. 2008;5(3):e67.
3. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting Outcome after Traumatic Brain Injury: Development and International Validation of Prognostic Scores Based on Admission Characteristics. PLOS Medicine. 2008;5(8):e165.
4. Antman EM, Cohen M, Bernink PM, et al. The timi risk score for unstable angina/non–st elevation mi: A method for prognostication and therapeutic decision making. JAMA. 2000;284(7):835-42.
5. Sylvester RJ, van der Meijden AP, Oosterlinck W, Witjes JA, Bouffioux C, Denis L, et al. Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. Eur Urol. 2006;49(3):466-5; discussion 75-7.
6. Kengne AP, Patel A, Marre M, Travert F, Lievre M, Zoungas S, et al. Contemporary model for cardiovascular risk prediction in people with type 2 diabetes. Eur J Cardiovasc Prev Rehabil. 2011;18(3):393-8.
7. Hilkens NA, Algra A, Diener HC, Reitsma JB, Bath PM, Csiba L, et al. Predicting major bleeding in patients with noncardioembolic stroke on antiplatelets: S2TOP-BLEED. Neurology. 2017;89(9):936-43.
8. Bolland MJ, Siu AT, Mason BH, Horne AM, Ames RW, Grey AB, et al. Evaluation of the FRAX and Garvan fracture risk calculators in older women. Journal of Bone and Mineral Research. 2011;26(2):420-7.
9. Hoster E, Klapper W, Hermine O, Kluin-Nelemans HC, Walewski J, Van Hoof A, et al. Confirmation of the Mantle-Cell Lymphoma International Prognostic Index in Randomized Trials of the European Mantle-Cell Lymphoma Network. Journal of Clinical Oncology. 2014;32(13):1338-46.
10. Wessler BS, Ruthazer R, Udelson JE, Gheorghiade M, Zannad F, Maggioni A, et al. Regional Validation and Recalibration of Clinical Predictive Models for Patients With Acute Heart Failure. Journal of the American Heart Association. 2017;6(11).
11. Sotomi Y, Cavalcante R, van Klaveren D, Ahn J-M, Lee CW, de Winter RJ, et al. Individual Long-Term Mortality Prediction Following Either Coronary Stenting or Bypass Surgery in Patients With Multivessel and/or Unprotected Left Main Disease: An External Validation of the SYNTAX Score II Model in the 1,480 Patients of the BEST and PRECOMBAT Randomized Controlled Trials. JACC: Cardiovascular Interventions. 2016;9(15):1564-72.
12. Takwoingi Y, Botello AP, Burr JM, Azuara-Blanco A, Garway-Heath DF, Lemij HG, et al. External validation of the OHTS-EGPS model for predicting the 5-year risk of open-angle glaucoma in ocular hypertensives. British Journal of Ophthalmology. 2014;98(3):309.
13. Collins GS, Mallett S, Omar O, Yu L-M**.** Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine. 2011;9:103-.

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 Medical Research Methodology. 2014;14:40-.
15. Little RJ, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Neaton JD, et al. The design and conduct of clinical trials to limit missing data. Statistics in Medicine. 2012;31(28):3433-43.
16. Baigent C, Harrell FE, Buyse M, Emberson JR, Altman DG**.** Ensuring trial validity by data quality assurance and diversification of monitoring methods. Clin Trials. 2008;5(1):49-55.
17. Collins GS, Reitsma JB, Altman DG, Moons KG**.** Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC medicine. 2015;13(1):1.
18. Bellary S, Krishnankutty B, Latha M**.** Basics of case report form designing in clinical research. Perspectives in Clinical Research. 2014;5(4):159-66.
19. Nahm ML, Pieper CF, Cunningham MM**.** Quantifying data quality for clinical trials using electronic data capture. PloS one. 2008;3(8):e3049.
20. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356.
21. Agniel D, Kohane IS, Weber GM**.** Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;361.
22. 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.
23. Pajouheshnia R, Peelen LM, Moons KGM, Reitsma JB, Groenwold RHH**.** Accounting for treatment use when validating a prognostic model: a simulation study. BMC Medical Research Methodology. 2017;17:103.
24. Sperrin M, Martin G, Van Staa T, Peek N, Buchan I**.** Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. arXiv preprint arXiv:1709.06859. 2017.
25. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine. 2015;12(3):e1001779.
26. Grady C, Eckstein L, Berkman B, Brock D, Cook-Deegan R, Fullerton SM, et al. Broad Consent For Research With Biological Samples: Workshop Conclusions. The American journal of bioethics : AJOB. 2015;15(9):34-42.
27. EUR-LEX. General data protection regulation. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC (accessed 16 May 2017).
28. Ohmann C, Banzi R, Canham S, Battaglia S, Matei M, Ariyo C, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. BMJ Open. 2017;7(12).
29. Gheorghe A, Roberts TE, Ives JC, Fletcher BR, Calvert M**.** Centre Selection for Clinical Trials and the Generalisability of Results: A Mixed Methods Study. PLoS ONE. 2013;8(2):e56560.

30. Rothwell PM. External validity of randomised controlled trials: “To whom do the results of this trial apply?&#x201d. The Lancet. 2005;365(9453):82-93.
31. Vergouwe Y, Moons KGM, Steyerberg EW. External Validity of Risk Models: Use of Benchmark Values to Disentangle a Case-Mix Effect From Incorrect Coefficients. American Journal of Epidemiology. 2010;172(8):971-80.
32. Zulman DM, Sussman JB, Chen X, Cigolle CT, Blaum CS, Hayward RA. Examining the Evidence: A Systematic Review of the Inclusion and Analysis of Older Adults in Randomized Controlled Trials. Journal of General Internal Medicine. 2011;26(7):783-90.
33. Shields KE, Lyerly AD. Exclusion of pregnant women from industry-sponsored clinical trials. Obstetrics & Gynecology. 2013;122(5):1077-81.
34. Hutchins LF, Unger JM, Crowley JJ, Coltman Jr CA, Albain KS. Underrepresentation of patients 65 years of age or older in cancer-treatment trials. New England Journal of Medicine. 1999;341(27):2061-7.
35. Lee PY, Alexander KP, Hammill BG, Pasquali SK, Peterson ED. Representation of elderly persons and women in published randomized trials of acute coronary syndromes. JAMA. 2001;286(6):708-13.
36. Schmidt AF, Groenwold RHH, van Delden JJM, van der Does Y, Klungel OH, Roes KCB, et al. Justification of exclusion criteria was underreported in a review of cardiovascular trials. Journal of Clinical Epidemiology. 2014;67(6):635-44.
37. van Staa T-P, Dyson L, McCann G, Padmanabhan S, Belatri R, Goldacre B, et al. The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. Health technology assessment. 2014;18(43):1-146.
38. Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Threats to Applicability of Randomised Trials: Exclusions and Selective Participation. Journal of Health Services Research & Policy. 1999;4(2):112-21.
39. Junghans C, Feder G, Hemingway H, Timmis A, Jones M. Recruiting patients to medical research: double blind randomised trial of "opt-in" versus "opt-out" strategies. BMJ. 2005;331(7522):940.
40. Khudyakov P, Gorfine M, Zucker D, Spiegelman D. The Impact of Covariate Measurement Error on Risk Prediction. Statistics in medicine. 2015;34(15):2353-67.
41. Pajouheshnia R, van Smeden M, Peelen LM, Groenwold RH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. (submitted).
42. McCarney R, Warner J, Iliffe S, van Haselen R, Griffin M, Fisher P. The Hawthorne Effect: a randomised, controlled trial. BMC Medical Research Methodology. 2007;7:30-.
43. McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects(). Journal of Clinical Epidemiology. 2014;67(3):267-77.
44. Wolfe F, Michaud K. The Hawthorne Effect, Sponsored Trials, and the Overestimation of Treatment Effectiveness. The Journal of Rheumatology. 2010;37(11):2216.
45. Sedgwick P, Greenwood N. Understanding the Hawthorne effect. BMJ : British Medical Journal. 2015;351.

46. Finniss DG, Kaptchuk TJ, Miller F, Benedetti F. Biological, clinical, and ethical advances of placebo effects. The Lancet. 2010;375(9715):686-95.
47. Grimshaw JM, Russell IT. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. Lancet. 1993;342(8883):1317-22.
48. Braunholtz DA, Edwards SJL, Lilford RJ. Are randomized clinical trials good for us (in the short term)? Evidence for a "trial effect". Journal of Clinical Epidemiology. 2001;54(3):217-24.
49. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353.
50. Fleming TR, DeMets DL. Surrogate end points in clinical trials: Are we being misled? Annals of Internal Medicine. 1996;125(7):605-13.
51. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Statistical Methods in Medical Research. 2014;26(2):796-808.
52. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Medical Research Methodology. 2016;16:163.
53. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Statistics in Medicine. 2016;35(2):214-26.
54. Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, et al. How to develop a more accurate risk prediction model when there are few events. BMJ : British Medical Journal. 2015;351.
55. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014;14:137.
56. Organization WH, Sciences CfIOoM. International ethical guidelines for health-related research involving humans: Geneva: Council for International Organizations of Medical Sciences; 2016.

Supplemental material

Supplement 1: Search query for trials (Pubmed).

The search was based on a search performed by Bastian et al.
(((Therapy/Narrow[filter]) AND (randomized trial OR control* trial))) AND ("2017/05/31"[Date - Publication] : "3000"[Date - Publication])

Supplemental reference

Bastian H, Glasziou P, Chalmers I (2010) Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLoS Med 7(9): e1000326. https://doi.org/10.1371/journal.pmed.1000326

CHAPTER 8

General discussion

It remains imperative that we elucidate factors that influence the performance of clinical prediction models. This will improve the design of prediction model studies and help with deciding whether a certain prediction model will make sufficiently accurate predictions for use in daily clinical practice. This thesis provides insight into the causes of heterogeneous or unexpected performance of prediction models and in turn provides recommendations for researchers when developing or externally validating prediction models. Specific attention was paid to the influence of treatments on the performance of prediction models.

Throughout the chapters of the thesis the following lessons were learned:

- There is considerable heterogeneity in the performance of prognostic prediction models for cardiovascular disease (CVD). A systematic review of three prognostic prediction models identified notable variation in the calibration and discrimination of the models across settings and a general trend for the models to apparently over-estimate CVD risk. Meta-regression analysis was unable to identify strong determinants of the between-study heterogeneity in model performance, but model discrimination was better in studies with greater variation in the characteristics of participants (Chapter 2).

- One possible cause of poor predictive performance is the use of treatments in the study cohorts used for model development or external validation, particularly when the intended use of the prognostic prediction model is to guide the decision to initiate those specific treatments. Studies that have developed or externally validated prediction models for CVD risk prediction have, by-and-large, not taken treatment use into account (Chapter 3).

- Failure to account for the use of "guided" treatments in studies that externally validate a prognostic model can lead to biased estimates of a model's discriminative ability and calibration. While excluding treated individuals from the validation study correctly removed the effects of randomly allocated treatments (as in a randomized trial (RCT)), inverse probability weighting (IPW) was preferred when treatments use was non-random (Chapter 4).

- Time-fixed methods to account for treatment use in a prognostic model study are not, in theory, suitable to account for treatment use that varies during follow-up. Although theoretically superior, advanced techniques such as the modelling of treatment use as a time-varying covariate in a prognostic model and the use of marginal structural models did not improve model performance in a case study (Chapter 5).

- An additional source of heterogeneity in prediction model performance is differences in the way predictors are measured across settings. When the magnitude and structure of the error with which a predictor is measured varies between settings (e.g. between development and external validation cohorts), the discriminative ability of the model can change (Chapter 6).

- The use of data from a RCT may provide an easier means to account for treatment use, as well as other benefits, such as high-quality, well-recorded measurements. However, factors including the selective inclusion of centres and patients and non-representative methods for predictor measurement may hamper the generalizability of prediction models developed using data from a RCT (Chapter 7).

Directions for future research based on the findings in this thesis

Future research should investigate how variation in the distribution of predictors and outcomes across settings affects the predictions provided by a model. This can be approached from three angles. First, the effect of variation in the distribution of patient characteristics (or "case-mix") from setting to setting can be further investigated. For example, recently advocated methods to evaluate the temporal and geographical variability of baseline risk need to be evaluated.[1] Also, better modelling of interactions between predictors may help in this. Second, the extent to which variation in the definition or measurement of predictors and outcomes across settings actually affects the performance of a prediction model needs to be evaluated by empirical and simulation studies. Third, the impact of (variation in) undergone treatments across patients or settings on prediction model performance requires further investigation. For example, this thesis has mainly discussed explicit treatment use (e.g. medication), while the effect of more implicit interventions during a prognostic model study such as lifestyle and behavioural changes have not been addressed. Finally, it is essential that future research considers these three different aspects together to assess their relative impact on prediction model performance. Existing approaches, such as benchmarking methods,[2,3] case-mix corrected performance measures [4] and approaches that draw comparisons across data samples [5-7] cannot yet distinguish between the above causes of variation across samples.

Towards better prognostic research: how to align aims and results

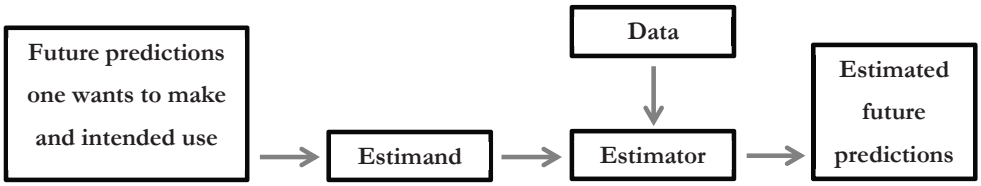
The issues addressed in this thesis and discussed above are related by a general theme, namely, a mismatch between what researchers aim to achieve when developing or externally validating a prediction model, and the actual results of such research. We

have seen how this can arise, for example through incorrect estimation of "untreated risks" in a prognostic study (using a data set in which individuals are (partly) treated). Inspired by the plea for estimands in randomised trials of therapeutic interventions, this thesis concludes with a proposal for a framework of so-called "prediction estimands", as an important first step to address these problems.

What are estimands, and how can they be used in prediction research?
In recent years, the concept of estimands has been advocated in the field of Phase III pharmaceutical trials, to increase the transparency of research and to harmonize trial objectives, analyses, and the results that are presented [8]. Succinctly, an estimand is the *intended estimate* of a study - a precisely defined theoretical construct, which can be used to inform the design of data collection and analysis of a study, primarily to yield research findings that are relevant to specific stakeholders (e.g. patients, care providers or guideline developers)[9-11]. It has been proposed that causal estimands (e.g. the estimands for a RCT) should comprise four components:[12] the target population, the patient endpoint (outcome), the specification of "intercurrent" events (i.e. post-baseline but pre-outcome, such as treatment switching or competing events), and a summary measure of the exposure or intervention effect. The estimand is selected based on the study objective and is subsequently used together with the available data to select a suitable statistical estimator, which provides an estimate, i.e. the final result of the study. In taking a structured approach, one aims to increase the likelihood of selecting the most suitable estimator and to obtain an estimate that is a quantification of the estimand that one is interested in. In this way the results of the study will be aligned with its aims.

Prediction model research may also benefit from adopting the concept of estimands. In this case the estimand of interest is not a summary measure of a causal relationship, but is instead a prediction for a future individual from the target population for which a prediction model is being developed or validated. In line with the PICOTS guidance for systematic reviews of prognostic model studies,[13,14] we could consider a prediction estimand to also consist of the proper target population, the relevant outcome (with an appropriate time horizon), consideration for intercurrent events (e.g. treatment initiation or competing events) and the necessary statistical measure (e.g. a predicted probability). A prediction estimand would also need to specify both a setting and a time-point at which the prediction will be made. As with causal estimands, we can use a prediction estimand to guide our selection of an appropriate estimator (e.g. regression modelling strategy), as well as evaluate how well the predictions made by the model align with our intentions. In addition, as prediction model development or validation is often conducted using data collected for other purposes, the prediction estimand can also be used to judge *a priori* the suitability of that data for developing or validating a prediction model (as discussed in Chapter 7). This is summarized in Figure 1.

Figure 1: An illustration of the path from estimand to estimate in prediction research



Example: defining an estimand for a prognostic model to predict CVD

As explained in Chapter 2 of this thesis, a prognostic prediction model can be used to select adults in primary care at high risk of CVD to initiate preventative interventions. Assume that one decides to develop a new prognostic prediction model. Given the future predictions one wants to make, the prognostic prediction model should provide predictions that represent the probability of a certain CVD outcome within, for example, 5-years, if individuals were to remain untreated over that time period. A full specification of that prediction estimand is presented in Table 1.

Table 1: An example prediction estimand for a study to develop a prognostic prediction model for CVD

Estimand characteristic	CVD example
Target population	General population; no history of CVD or current use of lipid lowering drugs.
Setting	General practice.
Timing	When meeting a general practitioner.
Intercurrent events	Preventative treatment: lifestyle changes and/or lipid or blood pressure lowering drugs. Non-CVD mortality.
Outcome	First major CV event, including a myocardial infarction, unstable angina or stroke within 5 years, if no preventative treatment is started.
Statistical measure	Predicted probability of this outcome.

Now that an estimand for this prognostic model development study has been defined, a suitable data sample can be obtained, either through the primary collection of new data or (more likely) the acquisition of a data set collected for other purposes. By pre-specifying an estimand before collecting or searching for the necessary data, we have a foundation to assess the suitability of the data for this prediction model study . As the estimand is a probability of CVD in the absence of treatment, one might consider collecting

data from an untreated population, such as from the control arm of a RCT evaluating the effectiveness of such treatment (see Chapter 7).[15] One also can try to acquire data that best matches our estimand in terms of the way that predictors and outcomes have been defined and measured. The estimand can then be used in combination with the chosen data to inform our selection of an estimator, in this case the statistical modelling techniques used to develop our prognostic prediction model.

The final, crucial step is to evaluate how well the estimates (i.e. predicted probabilities) provided by the developed prognostic model align with the estimand. Due to inadequacies in our data or statistical methods, we may find discrepancies between what we intended to predict and what the prognostic model actually predicts. For example, if 30% of individuals in the data began treatment during follow up and we suspect that it compromised the estimation of untreated risks, we might decide to modify our estimator to account for this (as described in Chapter 5). Having identified any such discrepancies between the estimand and estimates, we can readily report these alongside the prognostic prediction model and discuss how this may limit the use of the model for future patients (or certain patient sub-populations) - in line with existing recommendations [16] - to facilitate appropriate use of the model in clinical practice. The prediction model will still need to be evaluated in terms of its clinical usefulness and impact on healthcare, but by precisely specifying a prediction estimand we may improve the odds of achieving this.

Conclusions

From a pragmatic perspective, clinical research only holds value when the products of the research have a meaningful interpretation. Put differently, "'Useful clinical research' means that it can lead to a favorable change in decision making (when changes in benefits, harms, cost, and any other impact are considered) either by itself or when integrated with other studies and evidence in systematic reviews, meta-analyses, decision analyses, and guidelines". [17]

For prediction models to be clinically useful, the individuals using the model must understand precisely what the predictions provided by the model represent in order to determine the likelihood that the models will provide accurate and meaningful predictions. We have seen throughout this thesis that there are a number of methodological barriers, which complicate and often obscure the true interpretation of newly developed prediction models and as a consequence their ability to cause favourable changes in clinical decision making. Actions are needed to facilitate the selection of the best models for use in practice. We argue that improvements will be made if we precisely define and

report the aims and methods of prediction model research. We hope that by formally defining what it is we intend to estimate in future prediction model studies in terms of specific estimands and then reflect on the suitability of our estimators and any resulting discrepancies between our estimands and estimates, we will move towards prediction models that are transparent and benefit the care providers and individuals who will use them.

References

1. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Validation of prediction models: examining temporal and geographic stability of baseline risk and estimated covariate effects. Diagnostic and prognostic research. 2017;1:12-.
2. Vergouwe Y, Moons KGM, Steyerberg EW. External Validity of Risk Models: Use of Benchmark Values to Disentangle a Case-Mix Effect From Incorrect Coefficients. American Journal of Epidemiology. 2010;172(8):971-80.
3. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biometrical journal Biometrische Zeitschrift. 2008;50(4):457-79.
4. van Klaveren D, Gönen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings. Statistics in medicine. 2016;35(23):4136-52.
5. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. Journal of Clinical Epidemiology. 2015;68(3):279-89.
6. 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.
7. 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.
8. Mehrotra DV, Hemmings RJ, Russek-Cohen E, Group IEREW. Seeking harmony: estimands and sensitivity analyses for confirmatory clinical trials. Clinical Trials. 2016;13(4):456-8.
9. Little RJ, D'agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. New England Journal of Medicine. 2012;367(14):1355-60.
10. Akacha M, Bretz F, Ohlssen D, Rosenkranz G, Schmidli H. Estimands and their role in clinical trials. Statistics in Biopharmaceutical Research. 2017;9(3):268-71.
11. Akacha M, Bretz F, Ruberg S. Estimands in clinical trials–broadening the perspective. Statistics in medicine. 2017;36(1):5-19.
12. Committee IS. Final Concept Paper E9 (R1): Addendum to Statistical Principles for Clinical Trials on Choosing Appropriate Estimands and Defining Sensitivity Analyses in Clinical Trials dated 22 October 2014 Endorsed by the ICH Steering Committee on 23 October 2014. URL: http://www ich org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy E. 2014;9.
13. Moons KGM, de Groot JAH, 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 Medicine. 2014;11(10):e1001744.
14. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356.
15. 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.

16. Collins GS, Reitsma JB, Altman DG, Moons KM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement. Annals of internal medicine. 2015;162(1):55-63.
17. Ioannidis JPA. Why Most Clinical Research Is Not Useful. PLoS Medicine. 2016;13(6):e1002049.

Summary

Prediction models can be used to support clinicians when making diagnostic or prognostic assessments. By using information on an individual's demographic, genetic or clinical profile, for example, a prediction model can provide estimates of the individual's probability of having (diagnosis) or developing (prognosis) a certain outcome. These estimates provide additional information to health professionals and patients to guide or support decisions. For both diagnostic and prognostic prediction models to effectively support clinical practice, they must provide accurate, clinically relevant and interpretable predictions.

A number of factors that can affect the performance of a prediction model have been identified, particularly statistical aspects of the development of prediction models. At the same time, systematic reviews have found that many prediction models fail to provide consistently good predictions across populations or settings- a phenomenon exhibited even when prediction models have been developed using the appropriate methodology. Research into the heterogeneous performance of prediction models across settings has largely focused on differences in the distribution of patient characteristics (or "case-mix") across settings as a primary explanation. However, attention has been growing towards alternative explanations for the unexpected (often poor) performance of prediction models when externally validated or implemented in new settings.

This thesis examines potential sources of variation in the predictive performance of prognostic prediction models when applied in new individuals. Attention is primarily given to the issue of treatment use in studies that develop or externally validate prognostic models, how this can affect the accuracy and generalizability of such prediction models, and possible methodological solutions.

Chapter 2 of this thesis examines the performance of three cardiovascular disease (CVD) prognostic prediction models: the Framingham Risk Score, the Framingham ATP III model and the ACC/AHA Pooled Cohort Equations. A systematic review of studies that externally validated these three prognostic models was conducted. 1585 studies were identified, of which 38 (describing a total of 112 external validations) were eligible for inclusion in the review. Data were extracted on characteristics of the included studies and the reported performance of the three models, and studies were scored for their risk of bias. Following this, meta-analyses of two measures of performance (c-statistic and O:E ratio) were conducted. The c-statistic greatly varied across the studies and the three models generally overestimated CVD risk in the included studies. Finally, meta-regression analysis indicated that greater variation in patient characteristics and values of their clinical measurements may be associated with a larger c-statistic.

Chapter 3 elaborates on the potential issues that can be caused by treatment use by individuals in prognostic prediction model studies. A typology for treatments was proposed, to better understand when and how they should be taken into account. Following this, we conducted a systematic review of how treatment use has been addressed in studies that developed or validated cardiovascular risk prediction models. In total, 302 articles were included in the review, of which nearly one-third did not report any information on treatment use in their studies. Only one article reported information on treatment use during the study follow-up. Recommendations were provided to help improve the design, analysis and reporting of future studies to develop or validate prognostic models.

Chapter 4 specifically addresses the issue of treatment use in a study to externally validate a prognostic prediction model. The mechanisms underlying the potential effect of treatment use on the results of a validation study were examined in detail. Following this, four different approaches to account for unwanted treatment effects on measures of model performance were examined theoretically, and compared in a simulation study. An inverse probability weighting (IPW) approach was found to perform well when treatment was non-random.

Individuals in an observational study may begin using treatments at any point of follow-up and in many situations may stop and restart treatment over the follow-up period. Chapter 5 compares approaches to account for the effects of such time-varying treatment use when using observational data to develop a prognostic prediction model. Building on previous work that investigated methods for time-fixed treatments, we conducted a methodologic case study to investigate more sophisticated methods using real patient data. We compared seven approaches to develop a prognostic model to predict five-year mortality in individuals diagnosed with chronic obstructive pulmonary disease, accounting for the use of selective beta-blockers (SBBs) by individuals in the study. While the use of different approaches did alter the coefficients in the prognostic model and the risk predictions made by the model, this did not translate to a noticeable difference in model performance.

In chapter 6, a second source of heterogeneity in prognostic prediction model performance is investigated, namely, differences in the measurement of predictors across settings. When prediction models are developed, externally validated and implemented in different settings, the ways in which predictor variables in the model are defined and measured can vary. We argue that this could explain variation in prediction model performance across settings. To substantiate this claim, we formally defined differences in predictor measurements in terms of "measurement error". We then derived a general expression for the relationship between error in the measurement of a

continuous predictor and the area under the ROC curve (AUC) of a univariable logistic regression model. Subsequently, the effect of variation in error in the measurement of the diagnostic biomarker D-dimer on the AUC of a diagnostic prediction model was investigated, as an example. Through this, we were able to demonstrate that differences in the measurement of a predictor across settings can indeed explain variation in the performance of a prediction model.

Chapter 7 builds on the findings from previous chapters to provide recommendations for "best practices" when developing or validating prognostic models using data from randomized clinical trials (RCTs). Four key benefits that RCT data provide for prediction research were identified: 1) data completeness, 2) high quality of data, 3) randomized and well-characterized treatments and 4) detailed meta-data. Six issues to consider before using data from a RCT for prediction research were explored: 1) selectiveness of the trial participant population, 2) non-representative predictor measurements, 3) extraneous trial effects, 4) short-term and surrogate outcomes, 5) insufficiently large sample size and 6) trial participant consent for re-use of the data. Following this, we formulated guidance for researchers on how to appraise data from a RCT for its suitability for prediction research.

This thesis ends with a general discussion of some of the current challenges in prediction model research and the proposal of a general framework to address these challenges. First, the key methodological advances presented in the preceding chapters are summarized. Next, suggestions are given for directions for further research into methods for prediction modelling, including specific recommendations for "next steps" for research into the issue of treatment use in prediction model studies. Finally, we integrate the recommendations made throughout this thesis and propose a framework for defining and linking the aims, estimands and estimators of a prediction study. We suggest that by following this framework of "prediction estimands", researchers can improve the design and interpretation of future prediction model studies.

Samenvatting

Predictiemodellen kunnen worden gebruikt om clinici te ondersteunen bij het maken van diagnostische of prognostische beoordelingen. Door bijvoorbeeld informatie over het demografische, genetische of klinische profiel van een persoon te gebruiken, kan een predictiemodel schatten wat de waarschijnlijkheid van het individu is om een bepaalde uitkomst (diagnose) te hebben of wat voor ontwikkeling (prognose) ze door gaan maken. Deze schattingen bieden informatie aan medische professionals en patiënten om beslissingen te nemen of deze verder te ondersteunen. Om diagnostische en prognostische predictiemodellen de klinische praktijk effectief te doen ondersteunen, moeten ze accurate, klinisch relevante en interpreteerbare voorspellingen bieden.

Er zijn een aantal factoren geïdentificeerd die van invloed kunnen zijn op de prestaties van een predictiemodel, met name statistische aspecten in de ontwikkeling van deze predictiemodellen. Tevens hebben systematische reviews vastgesteld dat veel predictiemodellen er niet in slagen om consistent goede voorspellingen te geven in andere populaties of klinische omgevingen - een fenomeen dat zelf gebeurt wanneer predictiemodellen zijn ontwikkeld met behulp van de juiste methodologie. Onderzoek naar de heterogeniteit in prestaties van predictiemodellen in verschillende centra was grotendeels gericht op verschillen in de verdeling van patiëntkenmerken (of 'case-mix') tussen de centra als een primaire verklaring. Er is echter steeds meer aandacht voor alternatieve verklaringen voor de onverwacht slechteprestaties van predictiemodellen wanneer deze extern worden gevalideerd of geïmplementeerd in andere landen of centra.

Dit proefschrift onderzoekt potentiële oorzaken van variatie in de prestaties van prognostische predictiemodellen bij toepassing op nieuwe individuen. Er wordt vooral aandacht besteed aan de kwestie van het gebruik van behandeling in onderzoeken die prognostische modellen ontwikkelen of extern valideren, hoe dit de nauwkeurigheid en/of de generaliseerbaarheid van dergelijke predictiemodellen kan beïnvloeden en hoe dit methodologisch op te lossen valt.

Hoofdstuk 2 van dit proefschrift onderzoekt de prestaties van drie voorspellende modellen voor cardiovasculaire aandoeningen: de Framingham risk score, het Framingham ATP III model en de ACC/AHA Pooled Cohort Equations. Een systematische review van studies die extern deze drie prognostische modellen valideerden, werd uitgevoerd. 1585 studies werden geïdentificeerd, waarvan 38 (die in totaal 112 externe validaties beschrijven) in aanmerking kwamen voor deze review. Gegevens over kenmerken van de geïncludeerde studies en de gerapporteerde prestaties van de drie modellen werden genoteerd en studies werden gescoord op hun risico van bias (vertekening). Hierna werden meta-analyses uitgevoerd van twee statistieken die de prestatie van het model weergeven (c-statistiek en O:E-verhouding). De c-statistiek varieerde sterk tussen de onderzoeken en de drie modellen overschatten in het algemeen het risico op hart en

vaatziekten in de geïncludeerde studies. Ten slotte gaf meta-regressie-analyse aan dat een grotere variatie in patiëntkenmerken en hun klinische metingen werden geassocieerd met een hogere c-statistiek.

Hoofdstuk 3 gaat in op de mogelijke problemen die kunnen worden veroorzaakt wanneer individuen in prognostische predictiemodelstudies een behandeling gebruiken. Een kader voor behandelingen wordt voorgesteld om beter te begrijpen wanneer en hoe hier rekening mee moet worden gehouden. Hierna hebben we een systematische review uitgevoerd over hoe men in cardiovasculaire predictiemodel studies, zowel ontwikkeling als validatie, is omgegaan met het gebruiken van behandeling.. In totaal zijn 302 artikelen opgenomen in de beoordeling, waarvan bijna een derde geen informatie geeft over het gebruik van behandeling in hun studies. Slechts één artikel meldde informatie over het gebruik van de behandeling tijdens de follow-up van het onderzoek. Aanbevelingen werden gegeven om te helpen bij het ontwerp, de analyse en de rapportage van toekomstige studies om prognostische modellen te ontwikkelen of te valideren.

Hoofdstuk 4 gaat dieper in op het probleem van participanten die behandeling gebruiken in een onderzoek om een prognostisch predictiemodel zonder behandeling extern te valideren. De mechanismen die ten grondslag liggen aan het potentiële effect van het gebruik van de behandeling op de resultaten van een validatiestudie werden in detail bestudeerd. Hierna werden vier verschillende aanpakken theoretisch onderzocht om rekening te houden met ongewenste effecten op metingen van modelprestaties en vervolgens vergeleken in een simulatieonderzoek. De techniek inverse probability weighting (IPW)-bleek goed te presteren wanneer de behandeling niet-willekeurig was.

Individuen in een observationele studie kunnen op elk moment gedurende follow-up behandelingen gaan gebruiken en in veel situaties kan deze worden stopgezet en later opnieuw worden gestart. Hoofdstuk 5 vergelijkt aanpakken om rekening te houden met de effecten van tijd variërend behandelgebruik wanneer observationele gegevens worden gebruikt om een prognostisch predictiemodel te ontwikkelen. Voortbouwend op eerder werk dat methoden voor tijd gecorrigeerde behandelingen onderzocht, hebben we een methodologische casestudy uitgevoerd om meer geavanceerde methoden te onderzoeken met behulp van echte patiëntgegevens. We vergeleken zeven aanpakken om een prognostisch model te ontwikkelen voor het voorspellen van sterfte binnen vijf jaar bij personen met de diagnose chronische obstructieve longziekte, rekening houdend met het gebruik van selectieve bètablokkers . Hoewel het gebruik van verschillende benaderingen de coëfficiënten in het prognostische model en de risicovoorspellingen van het model veranderde, vertaalde dit zich niet naar een merkbaar verschil in de prestaties van de modellen.

In hoofdstuk 6 wordt een tweede oorzaak van heterogeniteit in de prestaties van (prognostische) predictiemodellen onderzocht, namelijk verschillen in hoe voorspellers gemeten worden. Wanneer predictiemodellen worden ontwikkeld, extern gevalideerd en geïmplementeerd in verschillende centra, kunnen de manieren waarop men voorspellende variabelen definieert en meet, aanzienlijk verschillen. We beargumenteren dat dit de variatie in de prestaties van predictiemodellen tussen centra kan verklaren. Om deze bewering te onderbouwen, hebben we formeel de verschillen in voorspellende metingen gedefinieerd in termen van een "meetfout". Vervolgens hebben we een algemene uitdrukking afgeleid voor de relatie tussen de meetfout van een continue voorspeller en het gebied onder de ROC-curve (AUC) van een univariabel logistisch regressiemodel. Hierna werd het effect van variatie in de meetfout van de diagnostische biomarker D-dimeer op de AUC van een diagnostisch predictiemodel onderzocht. Als gevolg hiervan konden we aantonen dat verschillen in de meting van een voorspeller in verschillende centra de variatie in de prestaties van een predictiemodel inderdaad kunnen verklaren.

Hoofdstuk 7 bouwt voort op de bevindingen uit eerdere hoofdstukken om aanbevelingen te doen voor "best practices" (beste aanpakken) bij het ontwikkelen of valideren van prognostische modellen met behulp van gegevens uit gerandomiseerde klinische studies (RCT's). Vier belangrijke voordelen die RCT-gegevens bieden voor voorspellingsonderzoek werden geïdentificeerd: 1) volledigheid van de gegevens, 2) hoge kwaliteit van gegevens, 3) gerandomiseerde en goed gedefinieerde, gestandaardiseerde behandelingen en 4) gedetailleerde metagegevens. Zes kwesties die in overweging moeten worden genomen voordat gegevens van een RCT voor voorspellingsonderzoek werden gebruikt, werden onderzocht: 1) selectiviteit van participanten aan de studie, 2) niet-representatieve metingen van voorspellende variabelen, 3) externe trial-effecten, 4) korte termijn- en surrogaatresultaten, 5) onvoldoende steekproefgrootte en 6) toestemming van participanten voor hergebruik van hun gegevens. Hierna hebben we richtlijnen geformuleerd die onderzoekers kunnen gebruiken om de geschiktheid van hun RCT voor een voorspellingsonderzoek te beoordelen.

Dit proefschrift eindigt met een algemene discussie over enkele hedendaagse uitdagingen in onderzoek naar predictiemodellen en een voorstel voor een algemeen kader om deze uitdagingen aan te pakken. Allereerst worden de belangrijkste methodologische ontwikkelingen die in de voorgaande hoofdstukken zijn gepresenteerd samengevat. Vervolgens worden suggesties gegeven voor verder onderzoek naar methoden voor het modelleren van voorspellingen, inclusief specifieke aanbevelingen aangaande de kwestie van het gebruik van behandeling in studies die predictiemodellen ontwikkelen of valideren. Ten slotte integreren we de aanbevelingen die in dit proefschrift zijn gemaakt en schetsen we een algemeen kader voor het definiëren en koppelen van de doelen en

de beoogde schatters van een voorspellingsonderzoek. We stellen voor dat onderzoekers door het volgen van dit kader, het ontwerp en de interpretatie van toekomstige voorspellingsonderzoeken kunnen verbeteren.

Dankwoord

To my dear supervisors, my sincerest thanks for your guidance, support and faith over the past three years.

Dear Professor Moons, Carl, I would like to thank you for giving the me opportunity to work with you and with the methodology group for the past three years. It has been a challenging journey, but we made it! You have always been incredibly open for discussion- a quality I hope I can apply in my own career. Finally I would like to thank you for the privilege of meeting and working with several excellent researchers, including the late, great Professor Doug Altman, who remains as someone I will look up to throughout my academic career.

Dear Dr Peelen, Linda, your support over these last few years has meant a huge amount to me. I'm so glad that you invited me to help with the EPIC-CVD project, to become one of the 'Three Musketeers' alongside Camille. Not only have I learned much from you about research and the field of clinical prediction models, but I also learned what it means to be a mentor, supervisor and a genuinely kind person. Our meetings with Noah were amongst my best memories of my PhD. Thank you.

Dear Dr Groenwold, Rolf, what a journey it's been these last five years. I owe so much of what I have learned over these years to your excellent tutelage. You have always been full of ideas and enthusiasm and you have given me so much motivation throughout the PhD. You have taught me much about what it means to be a researcher and work in academia and I will take those lessons forward in my career.

Dear Dr Reitsma, Hans, although you were not officially a member of my supervisory team, it has been an absolute pleasure working together with you and I have enjoyed discussing everything from prediction modelling to our vacation plans.

Dear Dr Van Smeden, Maarten, I'm glad that I was able to work with you these last two years. I admire your rigorous scientific approach and I hope I have been able to learn from it. I wish you all the success in Leiden.

Dear Dr Damen, Anneke, and dear Pauline, what a team we were! Even during some particularly grueling days of data extraction or difficult meetings to discuss our review papers, we worked together so well, and it was in these meetings where I learned much about the field of clinical prediction research. I want to thank you for your friendship as well as your hard work as colleagues.

To my fellow members of the Methodology group at the Julius Centre, it has been a privilege to work and learn alongside you over the last 5 years. I would especially like

to thank the members of Cochrane Netherlands, Lotty, Thomas, Rob, René for your support over the years.

Dear Noah, it has been a wonderful experience working together with you on your project. I could not have asked to supervise a more hardworking and diligent student. I wish you so much success with your PhD, and I hope one day we can work together again.

To my 'kamergenoten' in Stratenum and the Van Geuns, Kevin, Valentijn, Anne Karien, Carline, Faas, Sander, Katrien- thank you for all of the great times together. Dear, Timo it has been such a pleasure to get to know you better these last few years- I couldn't have asked for a better 'roomie'. Our converations about work and PhD life really helped to keep everything in perspective (and your knowledge of R was always super handy!). It was a real honour to be your paranymph and help you celebrate! Dear Giske, we have had a lot of fun over the past few years and I'm glad that you came over and joined Timo and me in the Van Geuns. I wish you all the best with the rest of your PhD, and most of all with your family.

To other friends around the Julius Centre, Josan, Marian, Ema, Mui, Rutger, thank you for everything. Abdel, bedankt voor de leuke gesprekken in het Nederlands! I also wish to thank the members of the biostatistics, Rebecca, Cas, René, Peter, from whom I've learned so much, and Renée Filius, for giving me the opportunity to work on many e-learning projects with Elevate. I would also like to thank Coby and Henk for all of the times you have helped me out over these years.

To my dear friends in the Netherlands, you have made these last few years unforgettable. Dear Mariana, your drawing of us at dim sum has sat on my desk these last two years and always made me smile. I look forward to many more years of friendship and conversations about food. Suus, it has been so much fun these last few years- I hope we can see more concerts together. To my international friends who I met in Utrecht, Quentin, Matevž, Ruben, Kostas, Mary, Matteo, SP, Maria, Lou, Ana, and others- I'm so glad we met over the past few years- your support and friendship during the PhD has helped so much.

My dear friends from England, all the gang from Devon, friends from Cambridge and London, thank you for everything. I want to especially thank to people who gave so much support via Skype- Laurence, Hannah, Dave, Erica, Alison and everyone else. Your support has meant everything to me. And to Laura, mon petit brie, thank you for the awesome cover design!

To my dear paranymphs, Loan and Rik,

Loan, I'm so glad that we met at the JC during my masters. You have been a brilliant friend. From squash twice a week, to climbing to exploring the world of veganism… I feel like you have been the saviour of my health! I'm looking forward to more trips to het Muzieklokaal and maybe one day talking to our favourite old couple. I'm really happy that you will be there to support me at the defence.

Rik, I don't have the words to express how much I value the support you have given me these past years. You have been a true friend and an excellent sparring partner for scientific discussion (I think I have learned nearly as much about epidemiology and statistics from you as I have craft beer!). You're going to smash the rest of your PhD.

Dear Anthony, this year could have been pretty overwhelming and yet it has been one of the best years of my life- thank you for so many amazing times and I'm looking forward to many more.

Finally, to my dear family, Dad, Mum, Soraya and Dave and of course Nan. This book is for you. You mean everything to me, and your unconditional support throughout this process has been something that I always knew I could fall back on. A thousand times, thank you.

List of publications

Publications based on studies presented in this thesis

Damen JA, **Pajouheshnia R**, Heus P, Moons KG, Reitsma JB, Scholten RJ, 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. Manuscript under revision.

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 Dec;1(1):15.

Pajouheshnia R, Peelen LM, Moons KG, Reitsma JB, Groenwold RH. Accounting for treatment use when validating a prognostic model: a simulation study. BMC medical research methodology. 2017 Dec;17(1):103.

Pajouheshnia R, Van Smeden M, Peelen LM, Groenwold RH. How variation in predictor
measurement affects the discriminative ability and transportability of a prediction model. Manuscript under revision.

Additional publications

Heus P, Damen JA, **Pajouheshnia R**, Scholten RJ, Reitsma JB, Collins GS, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. BMC medicine. 2018 Dec;16(1):120.

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. Journal of clinical epidemiology. 2016 Oct 1;78:90-100.

Pajouheshnia R, Pestman WR, Teerenstra S, Groenwold RH. A computational approach to compare regression modelling strategies in prediction research. BMC medical research methodology. 2016 Dec;16(1):107.

About the author

Romin Pajouheshnia was born on October 21st 1988, in Plymouth, United Kingdom. In 2010, Romin graduated from the University of Cambridge with a 2.1 Bachelor of Arts with Honors in Natural Sciences. During this time he specialized in immunology and virology, and conducted a 5-month research project at Addenbrookes Hospital, Cambridge. Following this, Romin spent 12 months teaching English as a foreign language in Ulsan, South Korea. In 2015, Romin graduated cum laude from Utrecht University with an MSc in Epidemiology, specializing in medical statistics. Under the supervision of Dr. Rolf Groenwold at the Julius Center, University Medical Center (UMC) Utrecht, he completed a 13-month research project in which he investigated an approach for comparing regression modelling strategies for prediction modelling. At this time he began to assist in analyses for the pan-European EPIC-CVD project. In addition, he co-developed one of the first Massive Open Online Courses provided by Utrecht University. The course "Clinical Epidemiology" has since seen over 14000 students enrolled. Thereafter, he began working on his PhD in epidemiology at the UMC Utrecht, under the supervision of Karel Moons, Linda Peelen and Rolf Groenwold. Throughout this period, Romin assisted in the teaching of several university courses at the UMC Utrecht and acted as daily supervisor of a Masters student. As of September 2018 Romin began working at the Department of Pharmaceutical Sciences, Utrecht University, as a postdoctoral researcher in methodology for pharmacoepidemiology.