Below you will find the comments from the reviewers, our response, and the changes we have made accordingly.

**Report from The BMJ’s manuscript committee meeting**

These comments are an attempt to summarise the discussions at the manuscript meeting. They are not an exact transcript.

Members of the committee were: Elizabeth Loder (Chair), Tim Cole (Statistics advisor), Jessamy Bagenal, José Merino, Georg Röggla, Tiago Villanueva, Wim Weber.

Decision: Put points

Detailed comments from the meeting:

Yours is an exhaustive systematic review of CVD prediction models. It makes a persuasive case that there are too many, often of poor quality. But readability of the review suffers somewhat from the same problem, with many models to describe. You have designated three classes: new models, validations of existing models, and extensions with new predictor(s). Reviewer 1 makes a good case to omit new predictor studies, as requirements are less than for other categories. This would focus the paper better and shorten it.

AUTHOR RESPONSE: We would like to thank you and the reviewers for reviewing our manuscript and for the positive and helpful comments regarding our manuscript. We have revised the manuscript taking into account all the comments, including omitting the incremental value studies to improve the readability of the review. Additionally, we have provided a point-by-point response to all the comments below. We believe these changes have strengthened the rationale and importance of our study.

Please change the title to reflect what research you did, it should not announce the findings.

AUTHOR RESPONSE: We have changed the title to "Prediction models for cardiovascular disease risk in the general population: a systematic review"

First, please revise your paper to respond to all of the comments by the reviewers. Their reports are available at the end of this letter, below.

In your response please provide, point by point, your replies to the comments made by the reviewers and the editors, explaining how you have dealt with them in the paper.

Comments from Reviewers

Reviewer: 1

Recommendation:

Comments:

A very thorough investigation that shows that the situation is much more chaotic than I had thought. However, I am not convinced that there is a way, given sufficient “head-to-head” comparisons, of finding one, or a few, superior risk-assessment method(s) from among those available; this seems to be the unspoken belief driving the paper. I guess a lot of people have their doubts about, eg QRISK2, despite its adoption in NICE guidelines, but at least it’s a rough-and-ready, intuitively correct, locally-acceptable approach (as ALL public-health measures can only be) to a national treatment protocol. (I am thus certainly of the authors’ opinion that no more time should be spent on refining existing models).

Is it worth publishing? Yes, if you want authoritative evidence for what most cardiologists (and a lot of patients) believe. No, if you hope it will lead to a better, tidier, more rational, etc, future.

AUTHOR RESPONSE: We thank the reviewer for this comment and we fully agree. We are similarly shocked by the current status of this field of CVD risk prediction. In the context of risk prediction, this is the most investigated and developed field, and, as the reviewer correctly remarks, the situation is very chaotic. It was this extreme diversity that we specifically wanted to highlight, and which we have studied in a systematic manner. Hence our main messages to the field are: stop developing yet another model; see how many models are already out there; see how few models have been actually validated even though they may be widely recommended; and see how few are compared head-to-head. This conveyor belt of continually developing new risk prediction models highlights more waste in research as most existing models are overlooked, as they are almost never validated or compared. Before contemplating developing another model, investigators should consider updating existing validated models. Finally, we agree with the reviewer that the current (chaotic) evidence is unclear on indicating which model to recommend in clinical practice. By conducting a systematic and comprehensive overview of the CVD risk prediction field and highlighting the inefficient nature of one model being developed after another with little examination of their utility, we believe it will lead to a better, tidier, more rational future of CVD risk prediction. Only by showing the extent of current chaotic status can we change our own and others’ behaviour.

Reviewer: 2

Recommendation:

Comments:

In their review of CVD prediction models, Damen et al. have undertaken an ambitious task of describing an extremely
heterogeneous field of research. Adding importance to the subject is the amount of resources invested in this area of science.

Although, the field of the article is important, the article in itself is of moderate importance. However, the data described in the article is interesting because it describes well the poor quality of most of the research articles in this field. Nevertheless, it probably will not help most readers of a general medical journal to make better decision as most of the suggestions derived from the conclusions are directed to those who design and supervise the general population based studies on which most risk prediction models are based on. Most often their work is limited, not by scientific aims or lack of vision, but by practical problems such as funding etc. The manuscript adds to existing knowledge by providing and up on the field but offers no big surprises.

AUTHOR RESPONSE: We thank the reviewer for this comment. Prompted by this comment we added a sentence to the Box "what this study adds". We also refer to our answer to reviewer 1 where we explain that we actually were surprised by the extreme diversity we found in this most developed field in the domain of risk prediction. The paper is meant to wake up ourselves and other researchers in the field of CVD risk prediction, and in this respect also practitioners, since they are (potential) users of these prediction models and often also develop clinical guidelines that may recommend these models. It is therefore important all parties understand the limitations of and differences between the huge number of existing CVD risk prediction models.

The research question seems clearly defined and well answered and the overall design of study is appropriate. The outcome measure is to provide descriptive data and mostly this is done adequately. The manuscript is well written and the methodology behind the literary research and the review itself seems sound.

The foremost problem of the manuscript is the heterogeneity of the subject it aims to describe. Mostly the conclusions that the authors propose are valid but they seem to be based on assumptions that all developed models (as defined by the authors) and incremental value assessments have been genuinely published for the purpose of competing with the pre-existing juggernauts dominating the field (for example SCORE, Framingham risk score and QRISK) or at least trying to offer a valid option for clinical work. In clinical practice, the existing validated models (SCORE, Framingham risk score and QRISK) are already implemented and no controversy exists of their replacement by new risk prediction models (with no validation).

AUTHOR RESPONSE: We thank the reviewer for this comment and we agree that this is indeed an assumption implicitly underlying our work. We do believe that this is not a strange assumption though. Why would so many researchers still be developing and publishing (up till this day and in major clinical journals) CVD risk prediction models in a field where such dominant models already exist and are recommended in guidelines? That would be because one believes that the current 'giants' are not of sufficient predictive performance or not applicable in their context. The aim of developing a new prediction model may not be to directly replace a 'giant', but rather to show whether better predictive performance can be obtained by their new model for their own clinical intent. But such conclusions require at least an external validation of the 'giants', which our review shows is hardly done. It is also not clear whether FRS, SCORE or QRISK are necessarily the 'best' models to use, as there is a paucity of head-to-head comparisons and alternative models may indeed be better in particular populations. In the UK for example, QRISK has only been compared against one version of Framingham. Unless we highlight the chaotic field of CVD risk prediction, new models will continue to be developed and published, with little evidence that they actually add anything to the already large and unyielding literature. We therefore believe this is the reason why our review is important for both researcher and practitioners involved in clinical guidelines. See also our response to the previous comment.

Major concerns:

1. Could the authors describe more accurately the articles (and their purpose) currently defined as "describing the development of new risk factor models"? One of the main concerns is that the models labelled "developed models" actually intended to compete with or outperform existing and well-established models (SCORE, FRS QRISK2)? If not, this could explain much of the shortcomings in the reporting described in the current report. As the authors present that only 46% the complete regression formula was published and only 22% were internally validated. These facts are reflected in the fact that only 64% of the developed models were externally validated. If these models are indeed developed to compete with pre-existing models in clinical use, the main conclusion of the current review could be simply that the methodology of most published studies (focusing on developing new risk models) should be improved.

AUTHOR RESPONSE: We thank the reviewer for this comment. Pro-
vably the added sentence to the Box "what this study adds" increases the clarity of the distinctions between developed models and existing models. We also refer to our answer to reviewer 1 where we explain that we actually were surprised by the extreme diversity we found in this most developed field in the domain of risk prediction. The paper is meant to wake up ourselves and other researchers in the field of CVD risk prediction, and in this respect also practitioners, since they are (potential) users of these prediction models and often also develop clinical guidelines that may recommend these models. It is therefore important all parties understand the limitations of and differences between the huge number of existing CVD risk prediction models.

2. The conclusion that "Future research should focus on assessing the validity and impact of existing models..." (abstract, page 5, line 130) seems to restricting. According to the manuscript, most developed models (66%) included the same set of major predictors (age, smoking, blood pressure, and blood cholesterol measurements) of CVD. It is reasonable to presume that examining the risk coefficients (effects) of these major risk factors in specific populations...
(if validated properly) can lead to more accurate prediction of CVD than using previous coefficients developed for generalization to all populations (within a continent as in SCORE) or developed in completely different ethnic population. Many variables (even the most common ones) used in risk prediction can vary in their definition between populations and they can also have different sized effects in different populations. The registry data maintained globally is highly heterogeneous in quality but in most countries with high accuracy of registry data, the applicability of such data can be very high for risk prediction. With constantly improving data management and electronic patient registries, could it be possible that specific population-level, easily available and monitored risk prediction models, can out-perform the old risk models based on fixed coefficients and baseline hazard?

AUTHOR RESPONSE: Thank you for the suggestion. We have paid little attention to possibilities of large data sets and registries, the tailoring of prediction models to these more local situations (rather than trying to develop a single super model), and the use of new techniques of model development. Prompted by this question we added the following to the discussion:

"Fourthly, more work is needed to evaluate the presence of heterogeneity in performance of different models across countries, allowing for tailoring of prediction models to different subpopulations. This can be achieved by combining the individual participant data (IPD) from multiple sources, including the increasingly available large registry datasets, and performing so-called IPD meta-analysis. Analysis of such combined or large datasets have the advantage of not only increased total sample size, but also of better addressing case mix effects, setting specific issues (e.g., inclusion of setting specific predictors), and better tailoring of existing models to different settings and consequently improving the robustness and thus generalizability of prediction models across subgroups and countries. Recently, prediction modelling methods for analysis of large, combined datasets have been proposed. After these efforts, generalizability of a developed and validated prediction model is still not good enough (e.g., because of too much differences between populations, treatment standards or data quality), more advanced methods for redevelopment of models can be used. Promising techniques are dynamic prediction modelling,60 61 modelling strategies that take into account treatment-covariate interactions,62 or other techniques such as machine learning,63 64 "

3. The authors have also searched for articles that describe the incremental value of a new predictor. The problem is that such a publication rarely even tries to specifically present new models for validation and/or to compete with existing validated models. However, in such case it is often a better option to use the best fit model in the background than try to implement pre-existing models with bad fit (due to any reason). Why not exclude these models from the current review?

AUTHOR RESPONSE: Thanks for this helpful suggestion. We agree with the reviewer and for clarity, we have now omitted the incremental value studies from the review.

4. The problem with including models that were used to calculate incremental value assessments of added risk markers is that they are often (due to practical reasons related to the publication demands) limited in their presentation. They distort the statistics describing the quality of reporting. For example, in abstract, the authors report (page 5, line 124): "... and for 92 (25%) crucial information was missing. Are these publications (of incremental value assessments) included in this figure presented in abstract? Currently in results section, these models (for incremental value assessment) are described only very briefly. I would recommend excluding these article from the review or not including them with summary statistics in abstract.

AUTHOR RESPONSE: The incremental value studies were indeed not included in the statistics describing the quality of reporting. Therefore, excluding the incremental value studies from the review, as suggested by the reviewer, has not changed these statistics.

5. As the authors concede, the literary research is already lagging behind. However, they justify the now 2.5 year interval between the literary search and current date by stating that:"including these articles is unlikely to change our main conclusions and recommendations". If the purpose of this review was to provide a comprehensive view on the subject, I feel an update would be in order. Most of the valid conclusions could be made even without the information provided by the current manuscript. In the present manuscript the authors cite the work of Beswick et al. (ref. 12) form the year 2008 (not from ten years ago as the authors claim). How have the conclusions changed since? If the last 2.5 years from June of 2013 to this day are not necessary for updating the message of the present manuscript, how has the field changed from 2008?

AUTHOR RESPONSE: The purpose of this review was to provide a systematic, transparent and extensive overview, not necessarily an up-to-date overview of all existing models for cardiovascular disease, because this would indeed be changing every day. As the reviewer suggests as well, there has been no large shift in science of CVD risk prediction during the last two years. Hence, our main conclusions -as we addressed above- are in our view still valid and would not change with the addition of these recent developed models. Moreover, it is the large number of newly identified articles in only two years, that actually underlines our main conclusions and reaffirms the necessity for changes regarding CVD risk prediction. The work of Beswick et al. is not published ten years ago, but they performed their search in 2003, which is more than ten years ago. Prompted by this comment we have changed the paragraph in the introduction as follows:

"However, the most comprehensive review was published performed by Beswick et al12 and includes models published more than 10 years ago (search carried out in 2003). More recent reviews have shown that the number of published prediction models has risen dramatically since then; and the most recent furthermore, these reviews have not systematically described the outcomes that the models intended to predict, the most frequent predictors, the predictive performance of all these models, and which developed prediction models are have been externally validated.13 14"

Also in the discussion we added the following paragraph on changes since the publication of Beswick et al.

"Since the publication of the review by Beswick et al in 2008, in which they searched the literature until 2003,
several major things have changed. The number of developed prediction models has tripled from 110 to 363, revealing problems such as the overwhelming number of prediction models, predictor definitions, outcome definitions, prediction horizons and study populations, and showing how poorly researchers make use of available evidence or existing models in the field. Although Beswick et al stated that "New prediction models should have multiple external validations in diverse populations with differing age ranges, ethnicity, sex and cardiovascular risk",12 we still found a great lack of validation studies for most developed CVD risk prediction models."

Minor issues:

1. How many developed models have incorporated the use of statins or other medications such as blood pressure lowering medication or family history of CVD.

AUTHOR RESPONSE: Modelling of treatment for cardiovascular diseases is indeed an important consideration. We have added the following information to the results:

"Treatment modalities were included in a minority of prediction models; 56 models (15%) included use of antihypertensive treatment and no models included use of lipid lowering drugs."

Jussi Hernesniemi
MD, PhD
University of Tampere

Reviewer: 3

Recommendation:

Comments:
Members of my family have heterozygous familial hypercholesterolemia (HeFH) and I have thought quite a bit about risk assessment for the primary prevention of cardiovascular disease. HeFH is a disease characterized by high risk of premature heart disease, but the risk is highly variable, with some untreated persons suffering myocardial infarction as early as their 20s or 30s and others not until late in life (e.g., 70s or older) or not at all. A number of studies have been done to try to discover the factors associated with premature heart disease in HeFH, but no model has been developed that explains more than a small part of the variance in risk and no validated risk calculator exists. As a result, most guidelines call for all HeFH patients to be treated starting in childhood, which results in overtreatment for some patients.

In addition, I am familiar with some of the risk calculators available for the general population and have used them myself to calculate my risk of an event (I do not have HeFH myself so am eligible to use most of these risk calculators). I have noticed that the various calculators vary with respect to the predictive variables that are included and the endpoints being assessed, as well as in the level of risk predicted. All in all, our ability to predict who will have a first cardiovascular event in the future is limited, meaning that any person wishing to lower their risk of such an event through interventions such as blood pressure and cholesterol lowering is rolling the dice in terms of whether they will benefit. I am also aware that many studies have been done in recent years on the effect of adding additional predictors to Framingham or other commonly used risk calculators and that the additional predictors usually only have a small additional effect, if any, over traditional risk factors alone. One possible exception is coronary artery calcium, but its use is this context is controversial. My impression has been that this unsatisfactory state of affairs is mainly due to the inherent unpredictability of the human cardiovascular system, rather than to a failure of research methods and reporting. For example, acute myocardial infarction results from sudden plaque rupture and clot formation, which partially or completely blocks the flow of blood through a coronary artery. Currently, our ability to predict the timing of such a sudden event is quite limited. Improvement in our ability to predict such events is would be very helpful to patients and their physicians as it would allow greater targeting of preventive interventions such as medications and lifestyle changes, and less under and overtreatment.

The authors of this systematic review have made an attempt to survey the field of prediction modeling for cardiovascular disease risk and outline some of the major shortcomings and difficulties in the field. They found quite a few, and make some specific suggestions in terms of improving methods of modeling, reporting, and research focus. For example, they suggest that the field would benefit from the formulation of a guidance with clear definitions of the relevant outcomes, predictors, and prediction horizons. I find myself unable to form a definite opinion on their suggestion that less emphasis be placed on the development of new CVD prediction models. In particular, my impression is that most validated models are based mainly on Caucasian and North American or European populations and that models are needed for the less developed countries and for non-Caucasian populations.

AUTHOR RESPONSE: We thank the reviewer for the comment and we fully agree there is a paucity of prediction models for less developed countries and for non-Caucasian populations. Prompted by this comment, we have added a statement to the results:

"One hundred and twenty-five articles described the development of 363 different models. The majority of prediction models (n=250, 69%) were developed using data from a longitudinal cohort study (Figure 3A); most originated from Europe (n=168, 46%) or the United States and Canada (n=132, 36%, Figure 3B). No models were developed using data from Africa. Several cohorts were used multiple times for model development, for example the Framingham cohort yielding 69 models in 23 papers."

Furthermore, we described this more explicitly in the discussion:

"We found much variability in geographical location of both model development and of model validation, but the majority of models were developed and validated in European and Northern American populations. Although the WHO states that more than three quarters of all CVD deaths occur in low- and middle income countries, 39 a prediction
In general, this paper seems mainly aimed at researchers in the field of cardiovascular risk modeling. As written, I do not think it will be useful to most patients or GPs.

AUTHOR RESPONSE: We thank the reviewer for this comment. Indeed it may seem that the paper was mainly aimed at researchers. However, as we refer also in our reply to reviewer 2, CVD risk prediction models are widely advocated in clinical guidelines across the globe, and used by medical practitioners. Moreover, they are increasingly approachable and used by individuals (e.g. patients) via web-tools or medical apps. Hence, we do believe that besides researchers, our inferences are also important for clinicians and patients. All involved parties need to acknowledge the limitations of and differences between the huge number of existing - and often directly available (via the web) - CVD risk prediction models. However, directed by this comment, we have now stressed this more clearly at several places in the text, including the "what this study adds"-box. See below.

To make it more user-friendly, I suggest that definitions of technical terms such as discrimination, calibration, and validation be provided.

AUTHOR RESPONSE: We thank the reviewer for this excellent suggestion and have added the following box:

Term Definition

Internal validation Testing a model's predictive accuracy by reusing (parts of) the dataset on which the model was developed. The aim of internal validation is to assess the overfit and correct for the resulting 'optimism' in the performance of the model. Examples are cross-validation and bootstrapping.22

External validation Testing a model's predictive accuracy in a population other than the development population.23

Prediction horizon Time frame for which the model is intended to predict the outcome.15

Discrimination Ability of the model to distinguish between people who do and do not develop the outcome of interest.24

Calibration Agreement between predicted and observed numbers of events.22

Updating Adjusting a previously developed model to a new setting or study population, to improve model fit in that population. Several forms of updating exist, including intercept recalibration, slope recalibration and refitting all coefficients of a model. 25 It is also possible to combine and update existing models.

In addition, summary points would be helpful.

We also followed this suggestion and restructured the "what this study adds"-box as follows:

What this study adds: There is an over-abundance of cardiovascular risk prediction models for the general population; yet few of these models have been externally validated, making them currently of unknown value for practitioners, policy makers and guideline developers. The majority of developed models are also insufficiently inadequately reported to allow external validation, let alone using them in practice.

At this point, researchers should may refrain from developing yet another similar CVD risk prediction model, and make better use of available evidence by validating, tailoring and making head-to-head comparisons, and tailoring of the promising existing models. Practitioners, clinical guideline developers, and patients should be aware that for many currently advocated CVD risk prediction models the predictive value is actually unknown.

A bit more explanation for some of their conclusions is needed as well. Parts of the paper read like an editorial but I found myself relatively unpersuaded by some of their points, such as the need to refrain from developing new models, but willing to listen if more reasons were given. It also would be helpful to know more about how and why the current state of affairs developed, e.g., file drawer effect, perverse incentives, funding priorities, etc.

AUTHOR RESPONSE: We thank the reviewer for the comment and agree that we could provide some more background discussion to this message. We therefore have added the following paragraph to the discussion to improve nuance:

"Presumably there are several reasons why researchers continue to develop a new CVD risk prediction model from scratch, such as the perceived lack of prediction models for their specific population (e.g. ethnic minorities) or specific outcomes (e.g. ischemic stroke), newly identified predictors, published articles reporting on bad performance of existing models in another setting, availability of data with higher quality (e.g. higher sample size, prospectively collected data), funding priorities, or merely self-serving to generate another publication. Nevertheless, our review clearly indicates that many of these studies are still similar in design and execution, as corresponding models often include the same (or similar) predictors, target the same (or similar) patient populations and predict the same (or similar) outcomes. Therefore, researchers are often -perhaps without knowing - reinventing the wheel and mostly introduce implicit knowledge when developing a prediction model from scratch. Given that there is a huge amount of literature on prediction of CVD outcomes for the general population, we think it is time to capitalise on prediction modelling research from scratch in this field. Over the past few decades, statistical methods for informatively building prediction models using established knowledge have substantially improved, and can be achieved by refining, updating, extending, and even combining the most promising existing models for prediction of CVD in the general..."
Reviewer: 4

Recommendation:

Comments:
Systematic review of prediction models for cardiovascular risk in the general population: too many models, too few validated and often insufficient reporting

Originality

Whilst Beswick and colleagues (Beswick AD, Brindle P, Fahey T, Ebrahim S. A Systematic Review of Risk Scoring Methods and Clinical Decision Aids Used in the Primary Prevention of Coronary Heart Disease (Supplement). London: Royal College of General Practitioners, 2008) have published a systematic review the conclusion from that report “The generalisability of risk scores in different populations cannot be assumed and there is currently little evidence that they are effective in improving health outcomes. Future research needs to be directed at refining the accuracy of prediction models and, most importantly, examining ways of turning them into effective clinical tools.” indicates that the objective and conclusions of the submitted paper are substantially different. I therefore believe it does add to the published literature and should be seen as original work.

By highlighting the plethora of prediction models and documenting their shortcomings, particularly around external validation, the authors provide much need guidance to researchers (and funders) on what should be the priorities for future studies. Secondly the results would be of interest to clinicians and policy makers with an interest in cardiovascular health who are potentially the end users of any developed models. They need to understand the limitations and differences between prediction models and be able to utilise ones appropriate for their populations.

Scientific reliability

The research objective is clearly defined.

Study design

The authors have undertaken a systematic review, albeit a descriptive one, and this is appropriate for the study question/objective.

Introduction

The introduction is brief, acknowledges previous published research and provides good justification for the current study.

Methods

The methods are clearly and adequately described. Although they have only searched from 2004 to 2013 I think their reliance on a previous review is acceptable given the nature of the topic. Having an end point is also important and they do acknowledge, when discussing the study limitations in the discussion (line 436), the large number of subsequent publications that have occurred after 2013. I agree with them that it is unlikely these publications would change their main conclusions or recommendations.

Their search strategy, screening and data extraction methods are consistent with other studies.

I agree that descriptive analysis was the best approach for the available data and current methods.

Results

The results were very interesting (surprising) even for a researcher/ and clinician who works in the field. For me there were several take home messages that I can use in my research, teaching and clinical work. Firstly the sheer number of studies, the fact that only a few have had extensive external validation, the wide variation in predictive performance, the inconsistency in outcome measures, the limited use of assessing the incremental value of predictors (despite the large number of predictors used) and the paucity of validated prediction models in some geographic areas.

Discussion

The discussion and conclusions are well written and consistent with the results. Although long I believe it would be hard to reduce its length and still do justice to the extensive results. The authors make sensible and logical recommendations that could potentially impact cardiovascular health care across the globe. I believe the messages are clear.

The limitations are acknowledged and discussed. It is important that the authors reference the limited evidence about the effect of risk prediction tools clinically and in particular on actual CVD incidence.

References
Different from the source cohort has been shown to have important implications for the prediction of cardiovascular disease (CVD) risk. This is because the wide use in practice of Framingham across diverse populations differs from the populations that had previously been neglected (for instance Africa/Asia/certain subgroups).

The differences between the models are not therefore just due to differences in the populations concerned. And indeed, deriving the Framingham equation and that used to support the derivation of QRISK2. One is based on relatively small population characteristics or setting differ in ways relevant to model adequacy; and 2) the data source itself might produce more research papers of course! But there might have been a case for new models either because 1) the paper comments on the lack of impact (Conclusions in the Abstract) we need to be clear that this systematic review included model development and validation studies, and studies of incremental variable inclusion, but not impact studies. It benefits from the more recent 2014 publication of the CHARMS checklist.

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

AUTHOR RESPONSE: We thank the reviewer for the question, which unfortunately is even at this point not yet easy to answer, notably because of the limited validations of existing models (except for the few well-known models in this field), and the poor reporting of the majority of the developed models. Nevertheless, prompted by this questions, we have added the following paragraph to the discussion:

"We also excluded articles describing models developed from or validated exclusively in specific diseased (patient) populations, e.g. patients with diabetes, with HIV-patients, with atrial fibrillation, or patients undergoing any surgery."

Finally, whilst I agree that the number of prediction models seems excessive, the authors reach this conclusion without very much discussion on why this is so and what the motivation is to produce yet more models (apart from producing more research papers of course!) But there might have been a case for new models either because 1) the population characteristics or setting differ in ways relevant to model adequacy; and 2) the data source itself might determine the suitability of the modelling approach. There is, after all, a huge difference between the data used to derive the Framingham equation and that used to support the derivation of QRISK2. One is based on relatively small but high quality prospectively collected cohort data and the other on routinely collected data from UK primary care. The differences between the models are not therefore just due to differences in the populations concerned. And because the wide use in practice of Framingham across diverse populations different from the source cohort has..."
attracted much criticism, it is perhaps not surprising that people have felt justified in modifying it regionally, or indeed in starting from scratch.

AUTHOR RESPONSE: We thank the reviewer for this suggestion. We have now added the following paragraph to the discussion:

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

I hope these comments are useful.
Tim Holt