29.09.2017

Dear José

Manuscript ID BMJ.2017.040627 entitled "Development and validation of QDiabetes®-2017 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study."

Thank you for reviewing our paper and giving us the opportunity to revise and resubmit it in the light of the reviewers' comments.

Committee comments

* The paper generated an interesting discussion at the manuscript meeting. We are aware that the original model is used widely by GPs in the UK and that a revision will have an impact on patient care. The reviewers raise many interesting concerns and we would like the authors to address each of these in turn. Please note that Richard Riley was the statistical consultant at the manuscript meeting and his statistical report is included in the comments from the reviewers. In addition, the editors had several questions that they would like the authors to address.

Authors response: thank you for the reviews and comment which we address in detail below.

* One editor with experience in primary care wrote that "It seems to make sense to update the risk prediction tool, but I am not sure how this will work in practice. While model A could be used to identify people at high risk of diabetes, then the next step would be to use a diagnostic test like FBG or HbA1c. So I am not sure what is the potential role of predictive models incorporating diagnostic tests [FBG and HbA1c (models B and C)]." Could you please discuss this issue in the manuscript?

Authors response: Model A can be used to identify people for diagnostic testing, but In people who don't have diabetes diagnosed at that point the blood test results can be used in combination with other risk factors in models B or C to give more accurate estimates of their risk of developing diabetes in the future. Use of model B in strategy 3 and model C in strategy 4 (figure 2) gives more accurate predictions of the future diabetes risk among those tested compared with strategy 1 or 2 based on blood tests alone. So, we have highlighted this more in the discussion (section 5.1).

* Another editor thought that a big part of the story, and perhaps an alternative interpretation, is that adding a lot of risk factors does not make a difference. "To me the interesting thing is that despite considering and in some cases adding in new risk factors (PCOS, antipsychotic use, gestational diabetes) the model performance doesn't improve all that much. This seems to be the old story of a few powerful risk factors dominating everything, so that there is a steep decline in the incremental value of additional risk factors. Maybe the authors and an editorialist should comment on that."

Authors response: For those individuals with these new risk factors, then the presence of these risk factors will substantially increase their absolute risk of diabetes. For an individual, this change in absolute risk could then push them over a threshold which may then result in different clinical management. However, the actual numbers of people in these particular groups are comparatively small so although there are slight increases in discrimination comparing model A with the current model (Table 5) the discrimination statistics are too crude to be able to detect these effects in small groups of individuals. Figure 3 included case studies where the impact of additional risk factors for an example patient would lead to different management. We have now updated the corresponding text in the results section to highlight this more explicitly and included additional text about this in section 5.1 of the discussion.

* One of the reviewers wrote to us after submitting his review to mention that "Tables 3 and 4 have incorrect titles – they present HRs for new diabetes but the current titles are: 'Adjusted hazard ratios (95% CI) for cardiovascular disease….' I think they mistakenly copied the titles from tables in their recently published QRISK3 paper."

Authors response: Thank you for spotting this. We have corrected the labels.

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.

Reviewer: 1 Laura Rosella. University of Toronto

Summary

This study presents the development and validation of the updated QRisk diabetes equations. These equations are widely used by GPs and integrated into primary care electronic records in England. Publishing updated prediction algorithms are an important to ensure clinical practice is using the most updated tools. Overall the study was comprehensive and well done. Some questions to follow:

Major comments

1. Please provide a sense of what proportions of English practices are using the EMIS computer system, and further for at least a year. I.e. the # that met inclusion criteria were reported but not the total number eligible based in all of England.
   Authors response: We have added the total number practices in England, the total using EMIS, the total contributing to QResearch and the total eligible for this study to sections 3.1 and 4.1.

2. Do patients actively de-register from a practice? What happens in the situation where someone doesn't come in for a while and perhaps has moved out but has not actively de-registered? Are they counted at-risk when in fact they are not being observed?
   Authors response: Patients deregister automatically as soon as they register with another practice since they can only be in one practice at a time so they will not be double counted.

3. Can the authors discuss the implications of (1) and (2) on the generalisability and performance of QRisk.
   Authors response: We have added some text to the discussion (section 5.3) to acknowledge the points about generalisability and the potential for misclassification bias of the outcome.

4. Please justify why statistical significance was used as a criteria for inclusion in the prediction model. A variable that is not statistically significant may still improve accuracy or discrimination. Generally speaking, p-values are not part of a predictive model building strategy.
   Authors response: There is a literature on developing prognostic models which does refer to selection of variables based on p values[1-3] including all of our own papers published in the BMJ. We used both hazard ratio and p value criteria in conjunction with clinical judgement to ensure that candidate variables and interaction terms are likely to be both clinically important and statistically relevant to reduce the possibility of

including weak or uninformative predictors leading to model over fitting and optimism[4]. We have added this justification to section 3.4.

5. From my understanding, Qrisk is used to identify people who should be further screened/tested (using FPG or AIC). However, the inclusion of FPG implies indeed they have been screened/tested, likely signifying the physician deems the patient to be at risk. Moreover, this affects the imputation. Those that that are missing FPG are probably at lower risk and thus their imputed FPG are likely overestimates (given that they are made to match the distribution of available data). Can the authors comment on this? Only 16% of the patients had complete data for FPG, smoking and BMI (it was unclear which variables were missing more often) – so the impact could be significant.
Authors response: the imputation approach incorporates the risk factors in the model, so imputed values for those with missing FPG are likely to be lower than the recorded values. We have added some text to section 5.3 about the percentage of patients with complete data. We have also included the complete case analysis for models A, B and C in supplementary tables 3 and 4. This shows hazard ratios were broadly similar in direction and magnitude to the analysis based on imputed data.

6. Do the % of data imputed differ between model A, B and C? This may affect the results as per comment 6. It would be helpful to include the % data missing/imputed by model and variables.
Authors response: We have used the same imputed dataset for deriving all three models so this didn't vary between the models. We have added a sentence in the methods section to clarify this (section 3.4).

7. Given the sample size – it would be feasible to estimate Models A,B and C based on complete data to see if the effect sizes meaningfully differ (i.e. you still have > 1million people with complete data). This would add some reassurance in light of the significant amount of imputation. Especially for key predictors (i.e. BMI) and potential for differential testing for FPG.
Authors response: We have added the results of the complete case analyses for models A B and C to the supplementary tables 3 to provide this reassurance. We have added the relevant text to the results (section 4.4) and discussion as suggested.

8. Some may question the inclusion of FPG such that it is circular to the outcome, given the cut-values for diabetes diagnosis. The authors may wish to comment on this.
Authors response: We have included FPG in model B as that is the recommended next step for patients whose initial diabetes risk is high based on model A and have explained this in section 5.1. The text now reads "Model A could be used to identify those at high risk of diabetes who require a fasting blood glucose test or HBA1C test. After identifying those patients who meet the criteria for a diagnosis of diabetes Model B could be used to refine the risk estimation in the remaining patients once the result of the fasting blood glucose test is known, since this information provides a more accurate assessment of risk".

9. Shouldn't Table 3 and 4 read "Adjusted hazard ratios (95% CI) for diabetes? Currently it says "cardiovascular disease"….
Authors response: thank you for spotting this. We have corrected this.

10. Given the comparison to existing models and implication for practice, the paper would benefit from the inclusion of reclassification and clinical utilization metrics, such as Net Reclassification Index or decision curve analysis.
Authors response: We have added decision curve analysis to the methods (section 3.7), results (section 4.7) and discussion (section 5.1). This shows model B had a slightly better net benefit than model C and both were better than model A. The graphs indicate that it is useful for risk thresholds up to about 40%.

11. The original model and simple model (A), now on the enhanced data (more practices and more complete data) actually performs quite well. This point could be emphasized.
Authors response: we have included some text in section 5.2.4 which compares the new models with the original models.

12. I wonder if the authors could add some thoughts on bias-variance trade-off – or over/under fitting. The more terms that are added to the model, the increase in complexity and the likelihood of overfitting. Ideally with prediction models, one is looking for that balance where the increased in complexity is equivalent to the reduction in variance, otherwise the models can perform worse. This is an explanation why in fact adding more variables does not outperform similar models. This also can explain why model C is not necessarily better than B.

Authors response: We have removed the sentence "there may be some over fitting of interaction terms given the number tested for inclusion" as this was added in error.
We have added the following text to section 5.3 of the discussion:

"There may be some over fitting but this is unlikely given the large number of events. Generally, it is recommended that there are at least 10 events per predictor variable, including the interaction terms to avoid overfitting[69]. In our most complex model (model C in women) there were 45 predictor variables. Our derivation sample had 178,314 events giving 3962 events per predictor variable which is nearly 400 times the minimum recommended level."

Minor comments

1. I was surprised how Model C (with the inclusion of A1C) significantly change South Asian ethnicity effects (Table 3). This may be worth commenting on with respect to the way risk is measured among South Asians.
Authors response: We have added an additional table (supplementary table 1a) which shows the distribution of risk factors by ethnic group in the derivation cohort and some corresponding text to the results (section 4.2). Testing for FBS and HBA1C was higher amongst south Asians compared with the white/not recorded group. South Asians had marginally higher HBA1C values which may partly explain this effect. We have highlighted the changes in hazard ratios between the different models by ethnic group in section 4.4 of the results and added a comment to section 5.2.4.

2. Figure 1 and 2 could be moved to the supplementary material.
Authors response: we have moved these to the supplement.

Reviewer: 2 Dan Lasserson University of Birmingham

1. Hippisley-Cox and Coupland present an update of the Qdiabetes prediction tool, derived and validated from UK electronic primary care data. As for the previous version of Qdiabetes, it is simple to implement in a UK primary care setting which is a considerable strength. The statistical methods are not controversial, and the inclusion of indicators of significant mental illness is to be applauded as such tools may go someway to reduce the mortality gap from physical disorders in this patient group.

   Authors response: thank you for these comments.

2. However, I think the definition of type 2 diabetes from primary care records needs some justification. I appreciate that the authors wish to maintain a consistent approach across the older and newer versions of QDiabetes, but there has been some methodological development to improve the accuracy of identification of patients with type 2 diabetes e.g. Clin Epidemiol. 2016 Oct 12;8:373-380 using THIN data. Given that the accurate identification of patients with type 2 diabetes is a critical outcome for derivation and validation I suggest that the authors either justify why their approach does not need alteration in line with the referenced method above or incorporate this in their analysis.
   Authors response: The method for classifying type of diabetes proposed by Sharma et al[5] was derived from THIN database and is based on records extracted from the Vision computer system. The Vision system is now only used by small minority of practices in England whereas EMIS is used by almost 60%. The THIN database has a non-standard coding system for "Additional Health Data" which is not present in QResearch so we couldn't apply this method to our database.

   As in other studies, we classified patients as having type 2 diabetes if they had a recorded diagnosis of diabetes and had not been prescribed insulin under the age of 35[6-8]. However, we excluded patients with either type 1 or type 2 diabetes from the study at baseline. Of the 377,053 patients with a diagnosis of type 1 or type 2 diabetes who were excluded in the derivation cohort, 34,195 (9.1%) were classified as having type 1 diabetes. This is very similar to other studies[9] including the one by Sharma et al where the proportion of patients with type 1 diabetes was 8.2%[5]. We have added this point to section 5.3 of the discussion.

Reviewer: 3 : Andrew Vickers

1. This is a very impressive study of a prediction model for diabetes. In developing the model, the authors have done all the right things, such as using a very large cohort of patients, separating validation and training cohorts, and following the TRIPOD statement. Unfortunately, the presentation of the results leaves a very large amount to be desired. What the authors do is throw a huge amount of numbers at the reader and somehow hope that they will make sense of them. For instance, in the key table 5 – the table I'd look to in order to determine whether the new model is better than the old model and whether e.g. it is worth taking a fasting glucose or not – the authors present no fewer than 72 different numbers. Which one should I be looking at to answer my question?

   Authors response: Table 5 presents 24 main statistics (with 95% CI). Comparing the new model B which includes fasting glucose with the new model A without fasting blood glucose

then this shows substantial increases in all of the discrimination statistics for men and women for model B (e.g. more than a 10% increase in variation explained) indicating that inclusion of fasting blood glucose does improve discrimination.

Supplementary table 5 provides 55 estimates for model A. How am I meant to derive from this table information on whether model A is worth using or not?

Authors response: We agree there is a lot of information in this supplementary table. We have provided this so the information is available for readers with a specific interest in how well the model performs in particular subgroups.

The same goes for the text of the results section, which is a rather rote repetition of prognostic metrics. For instance, the authors write: "In women, model A explained 50.5% of the variation in time to diagnosis of type 2 diabetes (R2), the D statistic was 2.07 and the Harrell's C statistic value was 0.834." They then give D, C and R2 for men for model A, women for model B, men for model B, women for model C and then men for model C, pretty much using exactly the same wording. After reading the paper, I'm left with a general feeling that the new model is probably a bit better than the old paper, and that using some additional tests can help. But I don't know whether the improvement with the new model is anything to write home about or whether it is worth doing the new tests. And I say this as a professional statistician who works with statistics like C every day. What the authors should do instead is to give statistics, and make logical, step-by-step arguments, that focus on clinical consequences. It is fine to mention discrimination and calibration, briefly, but then the authors need to move on to explaining what would happen to patients and populations if the models were used to make clinical decisions, such as whether to give intensive lifestyle advice or do more testing.

For instance, the authors might show a table that gives, for every 100,000 patients, the number of patients identified at high risk by a certain illustrative cut-point on the model, the number of diabetes cases identified early (i.e. sensitivity x prevalence) and the number of cases missed (i.e. [1-sensitivity] x prevalence). They might then compare these results for different cut-points or different models (e.g. original model vs. new model).

Authors response: Please note supplementary table 8 gives the values which are needed to do these calculations at different thresholds of risk. Figure 2 gives numbers and % of patients with diabetes identified early for different models and strategies.

Several of the prior Qrisk models have also published decision curves e.g. BMJ 2012;344:e4181 doi: 10.1136/bmj.e4181, which is a good way of evaluating clinical utility. I look at that curve and I fully understand how and in what circumstances the model is of value. In particular, the new model is better in the range 5% to 20%, which is exactly the sort of thresholds for decisions about cardiovascular risk prevention. I look at the current table 5, and I'm not so sure at all. Indeed, the BMJ has published a full explanatory paper describing the advantages of decision curves and similar approaches (BMJ 2016;352:i6 http://dx.doi.org/10.1136/bmj.i6).

Authors response: We have added decision curves for the new diabetes models (see response below).

Hence, the major suggestions are:

2. Dramatically reduce the number of measures of discrimination that are reported. I would recommend dropping reference to R2, sensitivity, specificity and D. If you are not sure why, simply ask yourself the question: "what levels of sensitivity / specificity / D / R2 is "good enough" to warrant clinical use of a model?" The lack of a coherent answer to that question illustrates the problem.
   Authors response: The purpose of this paper is to present the derivation and validation results in sufficient detail to allow other parties (including other researchers, heath economists, policy makers and those who write clinical guidance) to evaluate it and develop their recommendations which will also need to take account of resources and other clinical priorities. Whilst we agree that we have presented a large amount of data, we are aware that third parties do tend to make use of this and if it's not presented, will contact us separately about it. We note that reviewer 6 (Richard Riley) was satisfied with the information presented.

3. Include clinical implications of using different models with a population standardized to a size of e.g. 1,000,000 or 100,000 (cf, supplementary table 5, where results are given for 2,629,940 patients). Report: How many patients subjected to further work-up / treatment unnecessarily? How many patients destined to get diabetes are identified for early intervention? How many of those patients missed?
   Authors response: We have presented figure 2, figure 4 and supplementary table 8 which between them include all the information needed to scale the results to country level, region, local health economy or general practice. We think this is preferable to an arbitrary size and limits the number of tables presented.

4. Calculate net benefit and consider a decision curve.
   Authors response: Authors response: We have added decision curve analysis to the methods (section 3.7), results (section 4.7) and discussion (section 5.1). This shows model B had a slightly better net benefit than model C and both were better than model A.

Other comments

5. The rationale for including a hazard ratio criterion of >1.1 or <.90 is unclear to this reviewer. Predictiveness depends on both the effect size and prevalence (compare, for instance, a predictor with an odds ratio of 1.2 that is found in 50% vs. 0.5% of the population).
   Authors response. See comment above in response to reviewer 1 point 4.

6. The funnel plots are invalid. The idea behind of a funnel plot is that under the null hypothesis of no association between variance and the estimate of interest, the distribution of points on the graph will follow a funnel shape. This is not the case when the estimate of interest is an AUC. When the central estimate of AUC is around 0.85, one would expect a long tail of low AUCs for centers with low sample size.
   Authors response: We have these included these plots as this allows an assessment of the variability between study practices as recommended by Collins et al[10] although we have changed the text and the legends on supplementary figures 2a-d so that they are no longer referred to as funnel plots.

7. Rounding is inconsistent in many tables, suggesting that markup code (e.g. putdocx in Stata) has not been used.

Authors response: The reviewer is correct that we haven't used the Stata command putdocx. We have, however, specifically considered how best to round each set of results taking account of the width of the confidence intervals and magnitude of the estimates in line with previous publications.

8. Various graphs present hazard ratios by a predictor such as age. These are pretty uninterpretable. Please give absolute risks instead.
   Authors response: The hazard ratios are adjusted for the other risk factors in the model and take account of these whereas the absolute risks depend on the levels of individual risks factors so could not be concisely summarised in one graph.

   We also note that reviewer 4 indicated these graphs were adequate so we have not changed them (they are also consistent with many of our previous publications and so allow comparisons to be made between different models).

9. The calibration plot is non-standard, with the x axis being centile of predicted risk rather than absolute predicted risk. (If you want an ado file to do a standard calibration plot in Stata, please feel free to get in touch!)
   Authors response: we have changed the calibration plots as suggested (see figures 1a-c)

10. On a side note, the interface for the Qrisk calculator does not follow many of the principles for risk communication, such as avoiding technical terms (e.g. "your score has been calculated using estimated data") and including both probability of the event and probability of not having the event.
    Authors response: thank you for these points, which we will bear in mind for future versions of the web calculator and will discuss with our patient representatives.

Reviewer: 4 Christian. Göbl, Medical University of Vienna, Austria

1. In this prospective cohort study Hippisley-Cox and Coupland aimed to develop and validate an updated QDiabetes-2017 risk prediction algorithm to estimate the risk for type 2 diabetes. Therefore, the authors studies 8.19 Mio patients (178,314 incident cases) between 25 and 85 years (derivation cohort to build the model). In addition, a validation cohort containing 2.63 Mio patients (63,326 incident cases of type 2 diabetes) was also included in this study. The follow-up period was 10 years and risk estimation models were assessed for men and women separately. Thereby traditional risk factors (already included in the previous QDiabetes model) in addition to new risk factors like atypical antipsychotics and psychiatric disorders, use of statins, history of gestational diabetes mellitus (GDM) or polycystic ovary syndrome (PCOS) were included. Moreover, the additional predictive value of laboratory assessments (HbA1c and fasting glucose) were examined. Three different models were created for both sexes, showing adequate predictive accuracy: Model A (anamnestic variables); Model B (anamnestic variables + fasting glucose); Model C (anamnestic variables + HbA1c). Model B showed the best performance for predicting 10 year risk of type 2 diabetes with C = 0.90 and 0.88 for women and men, respectively.

2. This is an interesting and timely area of study and of particular importance for the general medical practitioner. I have following questions and suggestions:

   Authors response: thank you for this comment.

3. The authors have wide experience in modelling of epidemiologic data and I appreciate the use of modern statistical techniques like multivariate imputations of missing data and use of fractional polynomials. The functional form of the nonlinear relationships is adequately visualized (Figure 1), but please explain the function selection procedure for fractional polynomials with more detail (e.g. see W Sauerbrei, J. R. Statist. Soc. A. 2002). Model development (derivation cohort): Did the authors compare prediction errors of less and more complex models (e.g. by using resampling strategies like cross-validation). If so, I would suggest to include some of this data to ensure that overfitting was adequately addressed.

   Authors response: We selected second-degree fractional polynomial terms (FP2) and have added this to the methods. We used Akaike's Information Criteria to compare fit and performance of different models in the derivation cohort. We have added this to section 3.4. We didn't compare prediction errors of less and more complex models using cross validation.

4. As an advantage, the authors considered sex-specific risk factors and therefore included previous diagnosis of gestational diabetes or polycystic ovary syndrome into the risk estimation models. However, diagnostic criteria for GDM and PCOS have changed during the study period (patients were registered between Jan 2005 and Dec 2016). Also, the prevalence of GDM seems to be unexpectable low in this study (0.4%). Please comment on how these disorders were diagnosed and how changes in diagnostic criteria might influence the models. Please note that also diagnosis of type 2 diabetes has changed: In 2010 the ADA and 2011 the WHO consistently added HbA1c ≥6.5% to their diagnostic recommendations for overt type 2 diabetes and the ADA has recently updated their diagnostic procedures (ADA Standards of medical care 2017).

   Authors response: The definition of polycystic ovary syndrome and gestational diabetes were based on clinical computer diagnostic codes recorded in the patients' electronic health record. We have clarified this in section 3.3. We agree that the prevalence of gestational diabetes seems low and this is only partly explained by our reporting over ages 25-84 years rather than restricting to women of child bearing age. It is possible that some women with gestational diabetes were not diagnosed or that the diagnosis was not recorded. We have added a comment on the under-ascertainment of these variables to the discussion section 5.3.

5. How was fasting glucose examined (venous sample or capillary measurement)?

   Author: it was venous sampling. We have clarified this in the methods section 3.3

6. High risk for type 2 diabetes could be also examined by oral glucose tolerance testing and impaired glucose tolerance (another "prediabetic" condition) is diagnosed if 2h glucose exceeds 7.8 mmol/l. While this test is more time- and cost-intensive it has superior performance over fasting glucose or HbA1c to identify patients with high risk. Thus, I would suggest to include OGTT data if available or address this point in the discussion.

   Authors response: We haven't got OGTT data and have noted this as a limitation in the discussion section 5.3.

7. Overall, I think that this is a very interesting and timely manuscript.

   Authors response: thank you for your helpful comments and suggestions.

Reviewer: 5 Rod Jackson

This is an update of the QRISK group's 2009 BMJ publication of a Diabetes prediction algorithm and follows the same basic methodological approach. The group also published QRISK3 in the BMJ this year, following a similar modelling approach and which I also reviewed for the journal. So overall, I have no concerns about the general methodology, which has been widely peer reviewed.

Nevertheless, I have several issues to raise in regard to this particular paper.

1. Firstly, I am concerned about the validity of model B, which adds fasting blood glucose (FBG) to the main model (A), and model C, which adds HbA1C to Model A, because of the degree of missing data. By my calculations, only 14.5% of the derivation cohort had FBG measured and only 6.2% of the cohort had HbA1C measured. I am not an expert on multiple imputation, but I would like to recommend that the editors seek advice from an expert on imputation about the validity of imputing a variable when 93.8% (and 85.5%) of the cohort is missing the relevant variables. Not only is this extreme missingness but it is not missing at random.
   Authors response: thank you for these points – please see our reply to reviewer 1 points 5 and 7.

2. Secondly, I note that Fig 1g and 1h in the paper (HRs by age and BMI) look very different to the equivalent figures in the 2009 paper. Could the authors explain. Generally speaking HRs attenuate with age (as in the 2009 paper)> However in the paper under review, they increase until about age 65-70 years, and then decrease in older people.
   Authors response: The graphs presented in the original 2009 paper were different since we used a different comparator in presenting the hazard ratios. We have updated the plots for this paper to be consistent with the earlier paper and the patterns are more comparable. These can now be found in supplementary figure 1.

3. Thirdly, I would like to suggest that it is timely for the authors to start developing their models on the whole dataset rather than randomly splitting into derivation and validation datasets. I had planned to make this same suggestion wrt QRISK3 but forgot to. Risk prediction modelling has developed a lot in the last 10 years, in no small part because of the modelling work done by the QRISK group, and it is now accepted that there is no gain in randomly splitting such a huge dataset as the two two cohorts will always be identical. The extended TRIPOD statement (Moons et al Ann Intern Med. 2015;162:W1-W73) discuss this issue is some detail. Also Steyerberg, who is one of the authors, has written extensively on this topic and argues that when a dataset is large enough that you can split it, it is no longer necessary to do it. (see his 2009 book on Clinical Prediction models Chapter 17, page 302 "In sum, split sample validation is a classical but inefficient approach to model validation…….. Simulation studies have shown that rather large sample sizes are required to make split sample validation reasonable. But with large samples, the apparent validity is already a good indicator of model performance. Hence we may conclude that split-sample validation is a method that works when we don't need it."
   If the authors are keen on splitting the cohort, then it should be done by time or geographically and only as a sensitivity analysis, not the main analysis. However, I would suggest that they don't do it at all, because, as expected, none of their many published models have shown any benefit of splitting the cohort.

Authors response: We think this is an important point but are aware of different views on it (e.g. reviewers 3 and 6). We have therefore included the following text in section 5.3 "Some researchers argue that a split sample validation is not necessary in the circumstances where the sample is large enough to be able to split it[3], as in our study. Others, argue that a split sample validation is still valuable but it would be better to use a non-random selection of practices covering a broader range of settings or to split by geographical location or by time since randomly splitting a huge dataset is likely to result in similar populations[4]".

Reviewer: 6 Richard Riley

Thank you for the opportunity to review this paper. A very relevant topic and well-written article. It is important to update prediction models over time, as predictor effects and baseline risks may vary. Also statistical methodology improves, and so it gives the authors the chance to apply state of the art techniques. Upon review, I think this is generally a very good paper, as you would expect from these authors, using similar methods to recent papers such as QRISK3. It may be tempting to view this as another Q paper, but we must not forget that the models are being built on large, routinely available data from the UK, using sound statistical methods and rationale, and is clearly produced with much care and attention to detail. Therefore, it is an article that is both timely, well-motivated and clinically relevant, and would be welcome in the BMJ in my opinion.

Authors response: thank you for these comments

Focusing on statistical issues, I do have some important comments for the authors to consider in their revision. I am confident these are addressable by the authors and hope they help improve the article going forward.

1. Perhaps my major comment is that there is no consideration or adjustment for overfitting (shrinkage) upon model development. This means I am concerned that predictions may be slightly too extreme, due to overfitting to the observed data. Of course, the dataset is huge, and there are 178000 events, so I do not expect overfitting to be very big. Indeed, this is most likely the rationale for not doing it. However, it is an important part of model building (to check for overfitting) and the authors themselves say that: "There may be some over fitting of interaction terms given the number tested for inclusion" – so why not examine this statistically, for example using internal validation via bootstrapping to get the average calibration slope of a developed model in the bootstrap samples? (This is then the uniform shrinkage). This is perhaps some evidence of overfitting in the validation: "In women, the mean 10 year predicted risk was 3.62% for model A and 3.42% for Model B. The observed 10-year risk was 4.21% (95% CI 4.16% to 4.26%). In men, the mean 10 year predicted risk for model A was 4.97% and 4.71% for Model B. The observed 10-year risk was 5.56% (95% CI 5.48% to 5.61%)." - predicted risks are slightly too low then (although still very close). Would this have been improved if a shrinkage factor had been applied to the developed model after development?

   Authors response: Please see response to reviewer 1, point 12 on over fitting. In summary, we had added the sentence about 'overfitting of interaction terms' in error and have removed it. We don't think there is likely to be overfitting since the sample size is so enormous. We have also added the calibration slope overall and by subgroup in supplementary table 7 with some corresponding text to section 4.5.2 of the results. This shows the models were well calibrated overall and by age group with there was no evidence

of overfitting as the slopes were close to 1. Model B tended to have better calibration than model A within some of the subgroups.

2. The validation is done a new group of practices. I think this approach is reasonably ok. I see another reviewer suggests this is sub-optimal. I have some sympathy with that thought view, in that a random sample of omitted practices would be expected to perform well, as the random element will make the derivation and validation practices similar on average. If splitting is done, I would have preferred a non-random sample, specifically looking at practices that cover a broad set of case-mix and settings. But I think all this is a minor point. Indeed, if the response to part 1 is that there is very little overfitting, and that shrinkage was unnecessary, then I think all this is a mute point.
   Authors response: We have added some text to section 5.3 of the discussion on the merits of split sample approaches to validation. However, we have also done a validation within each general practice (see figure 2 in supplement).

3. The reporting of discrimination is well done, with a range of statistics given (one reviewer says to reduce these, but I don't agree), but calibration is not so well reported. In terms of examination of calibration performance, please can we see estimates of the calibration slope, alongside the overall agreement? Also, the calibration plots in Figure 3 should have CIs adding to each point, and actual risk values given on x-axis (not tenths).
   Authors response: We have added the CI to the calibration graphs now labelled figure 1a-c. The x-axis now gives the actual risk values.

4. Discrimination is examined in subgroups, but not calibration. Yet calibration is far more important, as discrimination is likely to reduce in subgroups by definition (as case-mix is narrowed), but calibration should be checked, using calibration slopes and overall calibration statistics.
   Authors response: we have added a new supplementary table 7 with the calibration slope overall and by subgroup and corresponding text to section 4.5.2 of the results.

5. I may have missed it, but crucially the actual developed model equations are not given in full. Some of the predictor effects are given in a table, but the fractional polynomial terms are not given (only in the figure). I also can't see the baseline survival values either at various time-points. To enable an independent validation, that does not require the user inputting values on the web tool, it is essential that the full models are given
   Authors response: information on estimates for the full models with baseline survival values is published on the qdiabetes.org website. We have added a statement to reinforce this to the paper (section 5.3). One of the other reviewers was able to access it on the website so wonder if there was an intermittent fault but it is working now.

6. How was fractional polynomial modelling (i.e. choice of non-linear trends) combined with multiple imputation and rubins rules? Specifically, the choice of trend may have been different in each imputation. Furthermore, it is known that the imputation should itself include the non-linear trends in order to be fully specified, as otherwise such trends may be lost or diluted. This is a methodologically taxing issue, but should be reported for transparency, even if a pragmatic approach was used to address things.

Authors response: We modelled the fractional polynomial terms using the complete data. We didn't include fractional polynomial terms in the multiple imputation. We then applied fractional polynomial terms to the imputed data before running the Cox models (section 3.4)

7. Methods says: "We used the regression coefficients for each variable from the final models as weights which we combined with the baseline survivor function evaluated" – how was the baseline survivor function estimated? Earlier on the authors refer to a non-parametric method, but please be clear that this was the approach used to get the S0(t) values for the final model.
   Authors response. Yes this was the approach we used. We have clarified this in the paper in section 3.4.

8. There are many version of the R-squared statistic. Please could you report which version you used? Specifically, it seems either Royston and Saurbrei's Rsquared_D or Royston's explain measure of variation R was used here, but which one?
   Authors response: We have updated section 3.5 to refer to the R^2 derived from Royston's D statistic.

9. STATA should be Stata.
   Authors response; We have changed this.

10. When comparing models, I think some consideration of the net-benefit function is essential (i.e. this is another of my most major comments). This links model predictions to clinical decisions, and compared the overall harms versus benefits of that decision, as described by Vickers et al. [1,2] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565-74. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6. doi: 10.1136/bmj.i6
    Authors response: We have added the decision curve analysis to the methods (section 3.7), results (section 4.7 and figure 4a and 4b) and discussion (section 5.1).

11. "For the new variables of interest in model A, atypical antipsychotics were associated with a 74% increased risk of type 2 diabetes in women …" – such %s should be accompanied by a CI
    Authors response: We have added the confidence intervals to the results section.

12. I wonder if there are competing risks, especially in the older age groups? If so, how was this handled, if at all? If it was not, then could risk predictions by too high in the older groups, due to the competing event of death being handled by censoring? See: Wolbers M, Koller MT, Witteman JCM, Steyerberg EW. Prognostic models with competing risks methods and application to coronary risk prediction. Epidemiology 2009;20:555-61
    Authors response: We have added the following text to section 5.3. "We haven't taken account of competing risks in this analysis since the results can be difficult to interpret and use in clinical practice[11]. However, not accounting for the competing risk of death in the elderly is likely to result in risk estimates which are too high in this age group".

13. Only 16% of patients had complete data for glucose measurements, smoking and body mass index. The authors rightly note this limitation. It is a major limitation, in my opinion. Therefore, Model B (the one shown to perform best) is based on a large amount of missing

data being imputed. I would note this limitation in the abstract itself. It suggests, at least to me, that further external validation is needed in more completely collected data, before we can be confident about using this model.
Authors response: We have added a sentence to the abstract as suggested and also emphasised the need for additional external validation in more complete datasets in the discussion of limitations (section 5.3). We have also added the results of complete case analyses for each model (see supplementary tables 3 and 4) and added comments on this in the results and discussion as indicated above.

14. Just a note that the C-statistic of 0.90 is enormous for Model B. Usually, for predictive models of future outcomes we struggle to get values over 0.70, so this is rather exciting to see such large value.
Authors response: yes we were pleased too!

15. I understand the computational demand, but I was surprised that only 5 imputations was used, as recent recommendations suggest more are needed. Usually the n of imputation is equal to the n% of patients with missing data for one or more variables, which here is about 84 I am guessing. But this may not be practical. Sterne et al say: "Report the number of imputed datasets that were created (Although five imputed datasets have been suggested to be sufficient on theoretical grounds, a larger number (at least 20) may be preferable to reduce sampling variability from the imputation process)" Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393. Therefore, given the large percentage of missing data combined with the relatively few imputations used, I think the authors need to give more reassurance that using n = 20 imputations would not chance the model importantly.
Authors response: We have added this to section 5.3 as a potential limitation although we were reassured by the complete case analysis. Our current server is simply not able to handle more than 5 imputations on such as large sample.

16. I tried the supplied username and Password to the Qdiabetes online tool but it did not work for me (i.e. I could not logon). Please can the authors check this for when I see the revised paper?

    Authors response: The login worked OK for the other reviewers and we have checked the website again and it's still working. Perhaps there was an intermittent fault. We will remove the need for a login prior to publication (the login is only there now to prevent people using it prematurely before the paper has been peer reviewed and published).

With best wishes and I look forward to reading the revision, Richard Riley

References

1. Royston P, Moons KG, Altman DG, et al. Prognosis and prognostic research: Developing a prognostic model. *BMJ* 2009;338(mar31_1):b604. doi: 10.1136/bmj.b604
2. Steyerberg EW, Eijkemans MJC, Harrell FE, et al. Prognostic Modeling with Logistic Regression Analysis. *Medical Decision Making* 2001;21(1):45-56. doi: 10.1177/0272989X0102100106

3. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer 2010.
4. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD StatementThe TRIPOD Statement. *Annals of Internal Medicine* 2015;162(1):55-63. doi: 10.7326/M14-0697
5. Sharma M, Petersen I, Nazareth I, et al. An algorithm for identification and classification of individuals with type 1 and type 2 diabetes mellitus in a large primary care database. *Clinical epidemiology* 2016;8:373-80. doi: 10.2147/clep.s113415 [published Online First: 2016/10/28]
6. Hippisley-Cox J, Pringle M. Prevalence, care, and outcomes for patients with diet-controlled diabetes in general practice: cross sectional survey. *Lancet* 2004;364(9432):423-8. doi: 10.1016/S0140-6736(04)16765-2
7. Hippisley-Cox J, Coupland C, Robson J, et al. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 2009;338:b880-. doi: 10.1136/bmj.b880
8. Roper NA, Bilous RW, Kelly WF, et al. Excess mortality in a population with diabetes and the impact of material deprivation: longitudinal, population based study. *BMJ* 2001;322(7299):1389-93.
9. Diabetes UK. Diabetes UK Facts and Statistics. London: Diabetes UK, 2016.
10. Riley RD, Ensor J, Snell KIE, 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 doi: 10.1136/bmj.i3140
11. Hippisley-Cox J, Coupland C, Robson J, et al. 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. [published Online First: 2010/12/15]