BMJ

BMJ 2012;345:e5900 doi: 10.1136/bmj.e5900 (Published 18 September 2012)

RESEARCH

Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study

Ali Abbasi *PhD fellow*[1 2 3], Linda M Peelen *assistant professor*[3], Eva Corpeleijn *assistant professor*[1], Yvonne T van der Schouw *professor of epidemiology of chronic diseases*[3], Ronald P Stolk *professor of clinical epidemiology*[1], Annemieke M W Spijkerman *research associate*[4], Daphne L van der A *research associate*[5], Karel G M Moons *professor of clinical epidemiology*[3], Gerjan Navis *professor of nephrology, internist-nephrologist*[2], Stephan J L Bakker *associate professor, internist-nephrologist/diabetologist*[2], Joline W J Beulens *assistant professor*[3]

[1]Department of Epidemiology, University of Groningen, University Medical Centre Groningen, Groningen, Netherlands; [2]Department of Internal Medicine, University of Groningen, University Medical Centre Groningen, Groningen; [3]Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, Netherlands; [4]Centre for Prevention and Health Services Research, National Institute for Public Health and the Environment (RIVM), Bilthoven, Netherlands; [5]Centre for Nutrition and Health, National Institute for Public Health and the Environment (RIVM), Bilthoven

Abstract

Objective To identify existing prediction models for the risk of development of type 2 diabetes and to externally validate them in a large independent cohort.

Data sources Systematic search of English, German, and Dutch literature in PubMed until February 2011 to identify prediction models for diabetes.

Design Performance of the models was assessed in terms of discrimination (C statistic) and calibration (calibration plots and Hosmer-Lemeshow test).The validation study was a prospective cohort study, with a case cohort study in a random subcohort.

Setting Models were applied to the Dutch cohort of the European Prospective Investigation into Cancer and Nutrition cohort study (EPIC-NL).

Participants 38 379 people aged 20-70 with no diabetes at baseline, 2506 of whom made up the random subcohort.

Outcome measure Incident type 2 diabetes.

Results The review identified 16 studies containing 25 prediction models. We considered 12 models as basic because they were based on variables that can be assessed non-invasively and 13 models as extended because they additionally included conventional biomarkers such as glucose concentration. During a median follow-up of 10.2 years there were 924 cases in the full EPIC-NL cohort and 79 in the random subcohort. The C statistic for the basic models ranged from 0.74 (95% confidence interval 0.73 to 0.75) to 0.84 (0.82 to 0.85) for risk at 7.5 years. For prediction models including biomarkers the C statistic ranged from 0.81 (0.80 to 0.83) to 0.93 (0.92 to 0.94). Most prediction models overestimated the observed risk of diabetes, particularly at higher observed risks. After adjustment for differences in incidence of diabetes, calibration improved considerably.

Conclusions Most basic prediction models can identify people at high risk of developing diabetes in a time frame of five to 10 years. Models including biomarkers classified cases slightly better than basic ones. Most models overestimated the actual risk of diabetes. Existing prediction models therefore perform well to identify those at high risk, but cannot sufficiently quantify actual risk of future diabetes.

Correspondence to: A Abbasi, Department of Epidemiology, University Medical Centre Groningen, Hanzeplein 1, PO Box 30.001, 9700 RB Groningen, Netherlands a.abbasi@umcg.nl

Extra material supplied by the author (see http://www.bmj.com/content/345/bmj.e5900?tab=related#webextra)

Appendix 1: Supplementary tables A-D
Appendix 2: Supplementary text
Appendix 3: References of excluded studies
Appendix 4: Supplementary figure A
Appendix 5: Supplementary figure B

Introduction

Type 2 diabetes is a large burden in healthcare worldwide.[1] Studies on lifestyle modifications and drug intervention have convincingly shown that these measures can prevent diabetes.[2 3] Early identification of populations at high risk for diabetes is therefore important for targeted prevention strategies and is necessary to enable proper efforts to be taken for prevention in the large number of individuals at high risk, while avoiding the burden of prevention and treatment for the even larger number of individuals at low risk, both for the individual and for society. The professional practice committee of the American Diabetes Association recommends screening for all overweight or obese adults (body mass index (BMI) ≥25) of any age who have one or more additional risk factors for diabetes such as family history or hypertension.[4] The European evidence based guidelines for the prevention of type 2 diabetes[5] and the International Diabetes Federation[6] recommend the use of a reliable, simple, and practical risk scoring system or questionnaire to identify people at high risk of future diabetes.

During the past two decades, many such prediction models have been developed.[7-11] Three recent reviews on this topic described existing prediction models and the predictive value of specific risk factors (such as metabolic syndrome) over a wide range of populations.[7-9] Surprisingly, however, the performance of less than a quarter of the prediction models was externally validated.[9-11] Because the performance of a prediction model is generally overestimated in the population in which it was developed, external validation of such models in an independent population, ideally by researchers not involved in the development of the models, is essential to broadly evaluate the performance and thus the potential utility of such models in different populations and settings.[12-15] Consequently, certain prediction models to identify those at high risk of diabetes cannot be recommended when external validity of available models is unknown.[12 16] Moreover, a direct comparison of the performance of the existing models in the same (external) validation cohort is essential to bridge the gap between the development of models and the conduct of studies for clinical utility.

The recent systematic reviews highlighted the need for an independent study to identify the existing prediction models and subsequently validate and compare their performance to support the current recommendations.[7-9] Few studies have externally validated such models, commonly not more than two or three at once, and almost always in medium sized cohorts.[10 11 14 17] We applied a more comprehensive approach as recently suggested.[14 15] Firstly, we carried out a systematic review to identify the most relevant existing models for predicting the future risk of type 2 diabetes. Then we used various analytical measures for validating[18] and comparing their predictive performance in a large independent general population based cohort—the Dutch cohort of the European Prospective Investigation into Cancer and Nutrition (EPIC-NL).[19]

Methods

Systematic literature search

We performed a systematic literature search according to the PRISMA guidelines,[20] when applicable. We searched PubMed for all published cohort studies that reported prediction models for the risk of type 2 diabetes until February 2011 using the following search string: ((“diabetes” OR “diabetes mellitus” OR “type 2 diabetes”) AND (“risk score” OR “prediction model” OR “predictive model” OR “predicting” OR “prediction rule” OR “risk assessment” OR “algorithm”)) NOT review [pt] AND English [LA]. We repeated this search for publications in German and Dutch. Finally, we checked systematic reviews and validation studies of prediction models to identify other relevant articles for our validation study. Because we did not perform a formal meta-analysis, the PRISMA items related to “protocol and registration” and “synthesis of results” for meta-analyses are not applicable to our study.

Studies were included if they met the following criteria: the study presented at least one formal prediction model or an update on a previously developed model; the endpoint was incident type 2 diabetes in a longitudinal design; and the population had to be at least partly white because the EPIC-NL cohort to be used for validation consists predominantly of white adults. We excluded studies using data on individuals with impaired glucose tolerance or impaired fasting glucose. Furthermore, we excluded models that used the two hour oral glucose tolerance test as a predictor variable because this was not available in our validation dataset and there was no reliable proxy variable available that could be taken as a substitute.

After review of the retrieved titles, two authors (AA and JWJB) independently reviewed the abstracts to select the relevant papers for full text review and subsequently reviewed and assessed the full papers. Discrepancies between the two reviewers were solved by having a third author (EC) review to reach consensus. For included studies, we made a primary plan to extract necessary data from the original studies to validate the models or contact the authors to obtain this information.

Table 1 summarises characteristics of the included studies⇓. The extracted data included the first author’s name, year of publication, country, name of study/score, number of cases and population, ascertainment of diabetes, duration of follow-up, statistical model, number of predictors, and reported performance of the model. The retrieved models were divided into models that contained only non-invasive predictors (“basic models”) and models that also included conventional biomarkers, such as glucose, HbA_{1c} lipids, uric acid, or γ-glutamyltransferase (“extended models”).

Validation cohort

The EPIC-NL cohort (n=40 011) includes the Monitoring Project on Risk Factors for Chronic Diseases (MORGEN-EPIC) and Prospect-EPIC cohorts, initiated between 1993 and 1997. The Prospect-EPIC cohort comprises 17 357 women aged 49-70 who participated in a breast cancer screening programme. The MORGEN cohort comprises 22 654 men and women aged 20-64 who were recruited through random population sampling in three Dutch towns (Amsterdam, Maastricht, and Doetinchem). At baseline, all participants were sent a general questionnaire and a food frequency questionnaire; these were returned when they visited the study centre for a medical examination. Reporting of the study results conforms to STROBE along with references to STROBE.[21]

We excluded 615 individuals with prevalent type 2 diabetes and 1017 with missing follow-up or who did not consent to linkage with disease registries. The 38 379 remaining participants were used to validate the basic models in a full cohort design. We applied similar exclusion criteria in a 6.5% baseline random sample (n=2604) in which measurements of conventional biomarkers were available,[19], leaving 2506 individuals. We used this random sample and all incident cases of type 2 diabetes to validate the extended models in a case cohort design.[22] Table A in appendix 1 provides baseline characteristics for the entire

cohort, the random sample, and the people with incident type 2 diabetes.

Assessment of predictor variables

Variables in the prediction models included in this study were assessed with a baseline general questionnaire for disease history and lifestyle variables. A validated food frequency questionnaire filled in at baseline was used to assess nutritional variables.[23] During the baseline visit, body weight, height, waist, and hip circumference, and blood pressure were measured and blood samples were drawn. Details of these procedures have been described elsewhere[19] and are shown in appendix 2.

Assessment of type 2 diabetes

Occurrence of diabetes during follow-up was self reported via two follow-up questionnaires at three to five year intervals in the MORGEN and Prospect cohort. In the Prospect cohort, incident cases of diabetes were also detected as glucosuria via a urinary glucose strip test, which was sent out with the first follow-up questionnaire. Diagnoses of diabetes were also obtained from the Dutch Center for Health Care Information, which holds a standardised computerised register of diagnoses at hospital discharge. Follow-up was complete up to 1 January 2006. Potential cases identified by these methods were verified against general practitioner (MORGEN and Prospect) or pharmacist records (Prospect only). Diabetes was defined as present when the diagnosis was confirmed by either of these methods. For 89% (n=1142) of participants with potential diabetes, verification information was available, and 72% (n=924) were verified as having type 2 diabetes and were included as cases of type 2 diabetes in this analysis.[24]

Data analysis

To evaluate the predictive performance of the retrieved prediction models, we used the original prediction models (regression coefficients with intercept or baseline hazard) as published. If the paper did not contain sufficient information, we asked the authors to provide us with the original model.[25 26] Particularly, we obtained regression coefficients[26] and the intercept of the model[25] by asking for complementary information. Using these original (regression) model formulas, we calculated the probability of developing type 2 diabetes per model for each individual in our study sample. Two authors (AA and JWJB) first matched the predictors of the original models with the variables available in our data. A direct match was available in our data for most variables. If a direct match was not possible, we replaced the original predictor with a proxy variable to avoid having to drop the model from our validation study. For example, we used non-fasting glucose values because fasting glucose values were not available in our data. Also, nutritional variables were collected with our food frequency questionnaire as continuous variables (g/day) and were re-coded into corresponding categories used in the prediction models by using Dutch portion sizes. Table B in appendix 1 provides an overview of the variables used in each of the prediction models, and appendix 2 gives the exact details on the proxy variables that were used.

We assessed performance of the models using measures of discrimination and calibration.[13] Discrimination describes the ability of the model to distinguish those at high risk of developing diabetes from those at low risk. The discrimination was examined by calculating Harrell's C (comparable with the area under the ROC curve), accounting for censored data.[27] Calibration indicates the ability of the model to correctly estimate the absolute risks and was examined by calibration plots. In a calibration plot, the predicted risk is plotted against the observed incidence of the outcome. Ideally the predicted risk equals the observed incidence throughout the entire risk spectrum and the calibration plot follows the 45° line. The calibration plot was extended to a "validation" plot as a summary tool.[18 27] Appendix 2 gives more details on information provided by this plot. Calibration was also tested with the Hosmer-Lemeshow goodness of fit statistic for time to event data.[18 27] Follow-up of our cohort was almost complete until about eight years: 3% were censored at 5 years, 5% at 7.5 years, and 44.6% at 10 years. To account for censoring when obtaining the observed probabilities for assessing calibration over, say, 10 years of follow-up, we first calculated for each individual the linear predictor and subsequently 10 year predicted outcome probability by the original survival models. This predicted probability was then divided into tenths, and we performed a Kaplan-Meier analysis per tenth, which accounts for the observed censoring. Per tenth, we obtained at the 10 year time point the observed outcome percentages, which in turn were compared with the 10 mean predicted outcome probabilities to obtain the calibration plot and measure of goodness of fit. This was done for each model and for the other time points (5 and 7.5 years).[28] Moreover, we reported the calibration slope for the logistic regression models[18] and calculated observed over predicted (expected) outcomes (O/E ratio) with 95% confidence intervals.[18 29] A ratio below 1.0 indicates overestimation of risk, and a ratio more than 1.0 indicates underestimation of risk.

Differences in the incidence of diabetes in our cohort and in the development populations led to significant deviation between observed risk in our cohort and predicted risk estimated by the prediction model. To reduce this source of miscalibration, we "recalibrated" each prediction model by adjusting the intercept (for logistic regression models) or the baseline survival function (for survival regression models).[28 30 31]

The original models were developed for different time periods of risk prediction (different "prediction horizons")—for instance, some models estimate 5 year risk and others 10 year risk. We therefore assessed the performance of each model for prediction of risk at 5, 7.5, and 10 years to account for the different time periods. For example, for 5 year risk, we considered individuals as incident cases if they had developed diabetes within the first five years of follow-up. Participants who developed diabetes after more than five years of follow-up were included in five year prediction as non-cases. A similar approach was followed for 7.5 and 10 year predictions. In addition, we performed a sensitivity analysis using the prediction horizon for which each model was developed in case this differed from 5, 7.5, or 10 years.

For the basic prediction models, which included only data from non-invasive clinical variables, we quantified their performance in the full dataset. The extended models were validated in the case cohort data. To account for this design, we applied an extrapolation approach that extends the case cohort data to the size of the full cohort.[22] This is achieved by extrapolating the non-cases of the random sample (that is, the total random sample of 2506 individuals minus 79 cases) to the number of non-cases in the full cohort (that is, the total sample of 38 379 individuals minus 924 cases). To do so, we substituted the non-cases in the full cohort (n=37 455) with a random multiplication of non-cases of the random sample (n=2427). On average, we multiplied the non-cases in the random sample by 15.4 (that is, 37 455 divided by 2427). Next, we merged the extrapolated data from non-cases to those from all the cases (total non-cases of 37 455 individuals plus 924 cases), recreating the size and composition of the full

cohort. In sensitivity analyses, we estimated the performance of the basic models in the re-sampled data from the case cohort and compared these results with those obtained from the full cohort. This allowed us to confidently use the extrapolation approach for the extended models in the case cohort design.

For most predictors data from less than 1% of the values were missing, although missing values occurred in 5% for family history of diabetes, about 15% for physical activity, and 20.5% for non-fasting glucose concentrations. Because an analysis of only the completely observed participants could lead to biased results,[32-34] we imputed these missing values using single imputation and predictive mean matching. As the percentage of missing values for the non-fasting glucose concentration was high, we repeated our analyses using only data from the MORGEN cohort, in which less than 10% of values for non-fasting glucose concentration were missing, as a sensitivity analysis. Table C in appendix 1 shows the number of missing values for all variables incorporated in the original model.

We carried out a third sensitivity analysis to account for the use of non-fasting glucose values, as we had to approximate the fasting glucose values included in the models by the non-fasting glucose values in our data. In this analysis, we excluded individuals with a non-fasting glucose of ≥11.1 mmol/L (n=130), as this cut point is considered as a high blood glucose concentration at which diabetes is suspected especially if it is accompanied by the classic symptoms of hyperglycaemia.[4] In another sensitivity analysis, we excluded 19 295 individuals (including 537 incident cases of diabetes) with fasting period of under two hours. In a fifth sensitivity analysis we excluded 255 individuals for whom we had no verification information of diabetes status.

All statistical analyses were conducted with SPSS version 18 (SPSS, Chicago, IL) and R version 2.11.0 (Vienna, Austria) for Windows (http://cran.r-project.org/).

Results

Systematic literature search

We scanned 7756 titles and selected 134 abstracts for review. Figure 1⇓ depicts the flow of the study selection process. We selected 46 articles for full text review and added six that were identified from other sources such as recent systematic reviews.[7-9] After full review of these 52 articles, we excluded 36 as they did not meet all inclusion criteria (appendix 3). The main reasons for exclusion were no prediction of the future risk of diabetes (n=15); validation study (n=10); no formal prediction models provided (n=6); and incomparable derivation populations (n=2) or unavailable data of predictors (n=3). Of three studies that used data from two hour oral glucose tolerance tests, we excluded two because they were cross sectional and one because it did not provide any prediction model.

Table 1 summarises the characteristics of the 16 studies included in this validation study.[25 26 35-48] Eleven studies described 34 basic models based on data that can be assessed non-invasively, including demographics, family history of diabetes, measures of obesity, diet, and lifestyle factors, blood pressure, and use of antihypertensive drugs. Of these 34 basic models, 12 models were presented as the final model.[8]

Nine studies described 42 extended models including data on one to three conventional biomarkers such as glucose, HbA_{1c}, lipids, uric acid, or γ-glutamyltransferase. Of these 42 extended models, 13 models were presented as the final model. The C statistics in the development datasets ranged from 0.71 for the Atherosclerosis Risk in Communities (ARIC) model to 0.86 for the FINDRISC full model. Only half of the studies reported measures of calibration, and almost all showed good calibration in the development datasets. Table B in appendix 1 shows the variables that are part of the prediction models.

Validation of prediction models

Table A in appendix 1 summarises the baseline characteristics of participants in the EPIC-NL study (for the full cohort, random sample, and incident cases of type 2 diabetes). During a median follow-up of 10.2 years (over 387 000 person years), we observed 924 incident cases (rate of 2.2 per 1000 person years). The observed 5, 7.5, and 10 year risks of incident diabetes were 1.3%, 1.8%, and 2.3%, respectively.

Tables 2⇓ and 3⇓ show the performance of the basic models and the extended models, respectively. The basic models performed well in terms of discrimination, with C statistics ranging from 0.74 (95% confidence interval 0.73 to 0.75) to 0.84 (0.82 to 0.85) for the prediction of risk of diabetes at 7.5 years. Similar but slightly higher C statistics were found for the 5 year risk prediction and slightly lower for the 10 year risk prediction of incident diabetes.

For the extended models, the discrimination was higher, with C statistics ranging from 0.81 (0.80 to 0.83) to 0.93 (0.92 to 0.94) for the risk at 7.5 years. Similar, but again slightly higher, C statistics were found for the 5 year risk prediction and slightly lower for the 10 year risk prediction of incident diabetes.

Both basic and extended models showed a poor calibration based on the Hosmer-Lemeshow test (P<0.001). Except for the EPIC-Norfolk and PROCAM models, all models overestimated the predicted against the observed 7.5 year risk of diabetes by 38.9% to more than 100%. Similarly, all observed to expected ratios were different from 1.0 (tables 2 and 3⇓⇓). The EPIC-Norfolk model underestimated the 7.5 year risk of incident diabetes by 73.9%. Figure A in appendix 4 shows the calibration plots for the original models.

After adjustment for differences in the incidence of diabetes between our cohort and the development populations, all prediction models showed better calibration (figs 2 and 3⇓⇓). For some of the models (such as the ARIC basic model) the calibration plot stayed close to the ideal line throughout the risk spectrum, whereas others showed severe overestimation, especially at higher predicted risks (such as Framingham continuous, DESIR, and BRHS models). Compared with the original models, the models adjusted for differences in the incidence of diabetes between the development and validation cohort performed better, with lower Hosmer-Lemeshow statistics, but deviation of calibration from ideal was still significant for all models, except for the KORA basic model (Hosmer-Lemeshow test P=0.17). For the KORA basic model, AUSDRISK, and EPIC-Norfolk model, calibration slopes were close to 1.0, but those were smaller or larger than 1.0 for other logistic regression models (tables 2 and 3⇓⇓). Figure B in appendix 5 shows the calibration plots including calibration statistics for each recalibrated models separately.

To further investigate the different effect size for each predictor, we compared hazard ratios for predictors between the validation cohort and one development cohort[49] as an example. We used data from the EPIC-Potsdam study[40] because the model was developed in the German cohort of EPIC using Cox proportional-hazards regression. Table C in appendix 1 presents the hazard ratios of the diabetes predictors incorporated in this risk score compared with those obtained in our validation cohort. The hazard ratios for age, intake of red meat, physical activity,

and current heavy smoking differed significantly (P<0.05) between both cohorts.

Sensitivity analyses

Tables 4⇓ and 5⇓ show the results of sensitivity analyses. Our results using the extrapolation approach for the case cohort design were similar when we looked at C statistics and Hosmer-Lemeshow statistics of 13 basic models obtained from the extrapolation approach compared with those from the full cohort design (for example, C statistics ranging from 0.74 (95% confidence interval 0.72 to 0.76) to 0.84 (0.82 to 0.86), and Hosmer-Lemeshow test P<0.001). Additionally, our results using data only from the MORGEN cohort with less than 10% missing values for non-fasting glucose were comparable with our results using both cohorts; C statistics ranged from 0.79 (0.76 to 0.81) to 0.92 (0.90 to 0.93) for 13 extended models. Exclusion of individuals with a non-fasting glucose of ≥11.1 mmol/L did not influence the results, both for the basic (C statistics ranged from 0.74 (0.72 to 0.75) to 0.83 (0.81 to 0.84)) and the extended models (C statistics ranged from 0.81 (0.80 to 0.83) to 0.93 (0.92 to 0.94)). Moreover, when we excluded the individuals with less than two hours' fasting or those without verified diabetes status, the C statistics were similar to those of the full cohort analysis. Finally, use of the prediction horizon for which the original models were developed hardly affected the results.

Discussion

An evaluation of the performance of 25 prediction models for type 2 diabetes in an independent Dutch cohort with over 10 years of follow-up showed that basic models perform similarly well in identifying individuals at high and low risk of developing diabetes. The performance was slightly better for extended models that included conventional biomarkers. With regard to the actual values of the predicted risks, all but two models overestimated the risk of developing diabetes, which improved slightly, but not sufficiently, after correction of the models for differences in incidence of diabetes between development and validation populations.

Strengths and limitations of study

All models were identified through a systematic literature search, and we included most existing prediction models in the validation study. Other strengths included the study's large sample size, prospective design, verification of incident diabetes, and extensive information on individuals' characteristics. Nevertheless, some limitations of our study need to be mentioned. Nearly all participants in the EPIC-NL cohort are white adults, and further studies are warranted to validate the models in other populations. In addition, the participation rate was about 40%.[19 50] We previously showed that such a low response rate might affect prevalence estimates of baseline characteristics of participants but does not cause bias in the examined associations.[50] We therefore consider that our cohort is appropriate for the purpose of our study. Although our data had certain limitations regarding availability of the variables, we made an effort to assign all variables and applied definitions as closely as possible. To handle missing variables, we performed single imputation and repeated the analysis in one of the two cohorts with lower missing values for glucose concentration, which gave similar results. It is therefore unlikely that these limitations influenced our results to a large extent. Next, we used data for non-fasting glucose concentration. We cannot rule out that this affected our results because glucose is an important predictor of diabetes. We therefore performed sensitivity analyses in which we excluded individuals with a non-fasting glucose of ≥11.1 mmol/L[4] and those who fasted for less than two hours, which again yielded similar results. This is in line with previous studies showing that using non-fasting lipid concentrations does not influence prediction of, for example, cardiovascular events.[51 52] Because we used data only from verified potential cases we could have missed false negative cases in the remainder of the cohort as type 2 diabetes can remain undiagnosed for several months to years. False negatives can lead to an underestimation of the C statistic as the linear predictor resulting from the predictor variables will be high, whereas their event status is that of a non-case. Given the large size of our cohort in combination with the low incidence of diabetes we do not expect this to largely change our findings. Similarly, false negative cases lead to underestimation of the observed risk in our cohort and this influences calibration. We adjusted for this effect, however, by correcting the intercept of the models to the incidence observed in our cohort. In addition, as the incidence is expected to be low,[53] potential false negative cases cannot account for the large overestimations of risk in the models observed in our study. Moreover, certain development cohorts used similar methods for verification of diabetes.

External validation of prediction models

The retrieved prediction models differed considerably in terms of type and number of predictors, age ranges, type of model, duration of follow-up, and outcome measure. Three recent systematic reviews presented overviews of studies that developed these models or validated some selected models.[7-9] These reviews, however, also indicated that most of these models were never validated in an external population. Our study has now evaluated performance of most developed prediction models for future diabetes in an external population and shows that most basic models perform well to identify those at high risk of diabetes and that extended models perform slightly better. Generally, the performance of a prediction model decreases when it is applied in a validation dataset. Despite this, our study showed that most of the basic models identified those at high absolute risk well, with C statistic over 0.80. This discrimination further improved for the extended models with C statistic of about 0.90. Surprisingly, the C statistics in our validation study were, in some cases, even higher than in their original development populations. This might be explained by differences in heterogeneity between the populations[30]: larger heterogeneity between individuals in a validation study can in some situations lead to a higher C statistic than in the development study. For example, variables like age, sex, and BMI might have larger heterogeneity in our study compared with the older population of the KORA study.[35] Although it would be of interest to explore whether performance of diabetes risk scores differs by age or sex, larger studies are warranted for these subgroup analyses. Another aspect that could influence model performance is the type of regression analysis used to derive the prediction model.[7] Most studies used logistic regression rather than survival models[7 8] and therefore do not account for censoring.[54] Similar to the results of the Framingham Offspring Study,[42] however, our results showed that the survival models do not necessarily perform better than the logistic ones.

Quantification of actual risk of future diabetes

All except two prediction models overestimated the absolute risk of diabetes in our validation dataset, which can partly be explained by the difference in incidence of diabetes between

development and validation populations. To account for this, we adjusted the models for difference in incidence, resulting in much better calibration. Significant deviations between the predicted and observed risks, however, remained for most models. There are various other explanations for the deviation in predicted versus observed risks. Firstly, in large cohorts the Hosmer-Lemeshow test is sensitive to small differences between the predicted and observed risks, so calibration can be indicated as significantly deviant by statistical tests even when the calibration plots indicate good calibration based on visual inspection and for practical purposes.[55] So, in large cohorts significant deviations on the Hosmer-Lemeshow test should be interpreted cautiously. Secondly, "mis"calibration can be caused by differences in how certain predictors, the outcome variable, or baseline characteristics of the study populations are measured, which can lead to different predictive effects.[13 30] For example, if the two hour oral glucose tolerance test is used to determine the presence of diabetes in a population, the incidence is likely to be higher, and among the cases there will be patients with a less severe form of the disease and different values for the potential predictors. This is also illustrated by comparing the effect sizes of the predictors of the German Diabetes Risk Score in our cohort, which showed significant differences for important predictors like age. It is important to note, however, that most prediction models showed overprediction, particularly at higher absolute risk. Some models might not have been well calibrated in the original populations.[7 9] Furthermore, the overestimation of risk at the higher end could be caused by overestimation of certain predictors in development populations with high risk individuals. Although it is important to accurately estimate the risk for people at high risk, it might not directly influence the effects of screening and public health strategies: interventions are often initiated beyond a certain threshold of absolute risk and overprediction beyond this threshold might therefore not necessarily lead to different treatment decisions. Certain models in our study, however, also overestimated the absolute risk in the lower ranges around 10%. Although decision thresholds for type 2 diabetes have not been determined, this prediction could be in the range of a threshold for a clinical decision. To use such models in clinical practice, calibration needs to be further improved.

Prior external validation of existing prediction models

Although the importance of external validation of prediction models is now widely acknowledged, only a quarter of existing prediction models have been externally validated, mostly in studies including only a single model and not reporting any measures of calibration.[7 9] To date, four studies have been published that performed a comparative external validation of several different models.[10 11 17 56] Two of these studies validated models for presence of diabetes rather than future risk of diabetes.[11 56] One prospective validation of three extended models[42 44 57] has been performed and showed C statistics ranging from 0.78 to 0.84 with underestimation or overestimation of the risk.[10] Another prospective validation study showed C statistics ranging from 0.74 to 0.90, without reporting calibration and performing adjustments.[17] These results are in line with the discrimination observed in our study. Altogether, the results from the previous reviews and our study suggest that most of the basic models performed similarly in terms of discrimination, whereas the Diabetes Population Risk Tool (DPoRT) showed slightly lower discrimination. The latter model was primarily developed to predict risk of diabetes at a population level, which could explain its slightly worse performance when it is applied on an individual level.[37]

Implications for use of prediction models in practice

Results from our study show that prediction models perform well to identify those at high risk of future diabetes, being a first prerequisite for use of such models in practice as currently recommended.[5 6] As expected[18 30] and observed in our results, however, the model should possibly be adapted to the local setting and purpose of the model and at least corrected for the incidence of diabetes of the population in which it is to be applied. The main relevance of prediction models is to correctly identify individuals at high risk, while avoiding the burden of treatment for individuals at low risk. This requires adequate discriminative power in the general population, as well as in populations characterised by a somewhat higher risk, such as those with excess weight. In public health practice, one would perhaps prefer to use a model including only a limited number of predictors based on non-invasive tests with the highest performance, which would favour use of a basic model. Noble et al[8] suggested seven models as most promising for use in clinical or public health practice, of which three were extended models (ARIC enhanced, Framingham, and San Antonio)[42 44 57] and four were basic models (AUSDRISK, QDScore, FINDRISC, and Cambridge Risk Score).[36 38 43 58 59] According to the current validation, it seems that this judgment is likely to be correct in statistical terms. The basic DESIR model that we additionally evaluated consisted of four predictors,[39] while most models—such as QDScore and AUSDRISK—consist of seven to 10 predictors. Interestingly, the models including only four to six predictors[35 39 43] performed similarly to the more extensive models.[36 38] We found that discrimination of two other basic models—KORA basic[35] and DESIR clinical equation[39]—which were not included in the list of Noble and colleagues,[8] approximated performance of the models incorporating more predictors. Moreover, the KORA basic model performed sufficiently to quantify absolute risk after recalibration. This suggests that a basic model like the KORA, which uses a limited set of non-invasive predictors, already provides good discrimination and good calibration and could therefore be useful in practice after appropriate adaptation of the model to the setting. The extended models including biomarkers could then perhaps be used only for those at high risk based on a basic prediction model. Finally, a model developed in one setting (such as public health data) or in a particular country does not necessarily need to be useful in another setting (such as secondary care) or country. As a next step, the utility of such models needs to be further investigated in clinical and public health practice.

Conclusions

Most of the basic prediction models including data on non-invasive variables performed well to identify those at high risk of developing type 2 diabetes in an independent population. The discriminative performance was slightly better for the extended models with additional data on conventional biomarkers. Most models, however, overestimated the actual risk of diabetes. Whether this influences treatment decisions needs to be further investigated. Hence, existing prediction models, even with only limited information, are valid tools to identify those at high risk but do not perform well enough to quantify the actual risk of future diabetes.

What is already known on this topic

There are many prediction models to estimate risk for future development of type 2 diabetes

An independent study to validate and compare the existing models is essential for assessing utility of prediction in practice, but has not yet been performed

What this study adds

Existing prediction models, even those that incorporate only four to six predictors, are valid tools to identify individuals at high risk for future development of type 2 diabetes

Actual risk for development of type 2 diabetes is generally overestimated, making it necessary to adapt models to local settings, and even then the accuracy of the estimated risk remains questionable

The impact of such prediction models on prevention or treatment decisions requires further investigation in clinical practice

We thank Statistics Netherlands and the PHARMO Institute for follow-up data on cancer, cardiovascular disease and vital status.

Contributors: AA, LMP, RPS, KGM, SJLB, and JWJB conceived and designed the study. AA, KGM, LMP, and JWJB analysed the data. AA, LMP, GN, and JWJB wrote the first draft of the manuscript. All authors contributed to the writing of the manuscript and agreed with manuscript results and conclusions. AA, LMP, and JWJB are guarantors.

Funding: This study was funded by the Netherlands Heart Foundation, the Dutch Diabetes Research Foundation and the Dutch Kidney Foundation, the Centre for Translational Molecular Medicine (project PREDICCt, grant 01C-104-07), Europe against Cancer Programme of the European Commission (SANCO), the Dutch Ministry of Health, the Dutch Cancer Society, the Netherlands Organization for Health Research and Development (ZonMW), and World Cancer Research Fund (WCRF), and the Netherlands Organization for Scientific Research project (9120.8004 and 918.10.615). None of the study sponsors had a role in the study design, data collection, analysis and interpretation, report writing, or the decision to submit the report for publication

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: The EPIC-NL cohort complies with the Declaration of Helsinki and was approved by the relevant local medical ethics committees. All participants gave written informed consent before study inclusion.

Data sharing: No additional data available.

1 Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care* 2004;27:1047-53.
2 Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 2002;346:393-403.
3 Group DPPR, Fowler SE, Hamman RF, Christophi CA, Hoffman HJ, Brenneman AT, et al. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. *Lancet* 2009;374:1677-86.
4 American Diabetes Association. Standards of medical care in diabetes—2011. *Diabetes Care* 2011;34(suppl 1):S11-61.
5 Paulweber B, Valensi P, Lindstrom J, Lalic NM, Greaves CJ, McKee M, et al. A European evidence-based guideline for the prevention of type 2 diabetes. *Horm Metab Res* 2010;42(suppl 1):S3-36.
6 Alberti KG, Zimmet P, Shaw J. International Diabetes Federation: a consensus on type 2 diabetes prevention. *Diabet Med* 2007;24:451-63.
7 Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med* 2011;9:103.
8 Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. *BMJ* 2011;343:d7163.
9 Buijsse B, Simmons RK, Griffin SJ, Schulze MB. Risk assessment tools for identifying individuals at risk of developing type 2 diabetes. *Epidemiol Rev* 2011;33:46-62.
10 Mann DM, Bertoni AG, Shimbo D, Carnethon MR, Chen H, Jenny NS, et al. Comparative validity of 3 diabetes mellitus risk prediction scoring models in a multiethnic US cohort: the Multi-Ethnic Study of Atherosclerosis. *Am J Epidemiol* 2010;171:980-8.
11 Lin JW, Chang YC, Li HY, Chien YF, Wu MY, Tsai RY, et al. Cross-sectional validation of diabetes risk scores for predicting diabetes, metabolic syndrome, and chronic kidney disease in Taiwanese. *Diabetes Care* 2009;32:2294-6.
12 Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ* 2009;338:b375.
13 Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:b605.
14 Collins GS, Moons KG. Comparing risk prediction models. *BMJ* 2012;344:e3186.
15 Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. *BMJ* 2012;344:e3318.
16 Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006;144:201-9.
17 Schmid R, Vollenweider P, Bastardot F, Waeber G, Marques-Vidal P. Validation of 7 type 2 diabetes mellitus risk scores in a population-based cohort: CoLaus study. *Arch Intern Med* 2012;172:188-9.
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;21:128-38.
19 Beulens JW, Monninkhof EM, Verschuren WM, van der Schouw YT, Smit J, Ocke MC, et al. Cohort profile: the EPIC-NL study. *Int J Epidemiol* 2010;39:1170-8.
20 Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;339:b2700.
21 Von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;335:806-8.
22 Volovics A, van den Brandt P. Methods for the analysis of case-cohort studies. *Biometric J* 1997;39:195-214.
23 Ocke MC, Bueno-de-Mesquita HB, Goddijn HE, Jansen A, Pols MA, van Staveren WA, et al. The Dutch EPIC food frequency questionnaire. I. Description of the questionnaire, and relative validity and reproducibility for food groups. *Int J Epidemiol* 1997;26(suppl 1):S37-48.
24 Sluijs I, van der A D, Beulens JW, Spijkerman AM, Ros MM, Grobbee DE, et al. Ascertainment and verification of diabetes in the EPIC-NL study. *Neth J Med* 2010;68:333-9.
25 Alssema M, Vistisen D, Heymans MW, Nijpels G, Glumer C, Zimmet PZ, et al. The Evaluation of Screening and Early Detection Strategies for Type 2 Diabetes and Impaired Glucose Tolerance (DETECT-2) update of the Finnish diabetes risk score for prediction of incident type 2 diabetes. *Diabetologia* 2011;54:1004-12.
26 Kahn HS, Cheng YJ, Thompson TJ, Imperatore G, Gregg EW. Two risk-scoring systems for predicting incident diabetes mellitus in US adults age 45 to 64 years. *Ann Intern Med* 2009;150:741-51.
27 Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361-87.
28 Van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Stat Med* 2000;19:3401-15.
29 Liddell FD. Simple exact analysis of the standardised mortality ratio. *J Epidemiol Community Health* 1984;38:85-8.
30 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;172:971-80.
31 Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;61:76-86.
32 Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393.
33 Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59:1087-91.
34 Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med Res Methodol* 2010;10:7.
35 Rathmann W, Kowall B, Heier M, Herder C, Holle R, Thorand B, et al. Prediction models for incident type 2 diabetes mellitus in the older population: KORA S4/F4 cohort study. *Diabet Med* 2010;27:1116-23.
36 Chen L, Magliano DJ, Balkau B, Colagiuri S, Zimmet PZ, Tonkin AM, et al. AUSDRISK: an Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. *Med J Aust* 2010;192:197-202.
37 Rosella LC, Manuel DG, Burchill C, Stukel TA. A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). *J Epidemiol Community Health* 2011;65:613-20.
38 Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 2009;338:b880.
39 Balkau B, Lange C, Fezeu L, Tichet J, de Lauzon-Guillain B, Czernichow S, et al. Predicting diabetes: clinical, biological, and genetic approaches: data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR). *Diabetes Care* 2008;31:2056-61.
40 Schulze MB, Hoffmann K, Boeing H, Linseisen J, Rohrmann S, Mohlig M, et al. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. *Diabetes Care* 2007;30:510-5.

41 Simmons RK, Harding AH, Wareham NJ, Griffin SJ. Do simple questions about diet and physical activity help to identify those at risk of type 2 diabetes? *Diabet Med* 2007;24:830-5.
42 Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB Sr. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. *Arch Intern Med* 2007;167:1068-74.
43 Lindstrom J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care* 2003;26:725-31.
44 Stern MP, Williams K, Haffner SM. Identification of persons at high risk for type 2 diabetes mellitus: do we need the oral glucose tolerance test? *Ann Intern Med* 2002;136:575-81.
45 Stern MP, Morales PA, Valdez RA, Monterrosa A, Haffner SM, Mitchell BD, et al. Predicting diabetes. Moving beyond impaired glucose tolerance. *Diabetes* 1993;42:706-14.
46 Wannamethee SG, Papacosta O, Whincup PH, Thomas MC, Carson C, Lawlor DA, et al. The potential for a two-stage diabetes risk algorithm combining non-laboratory-based scores with subsequent routine non-fasting blood tests: results from prospective studies in older men and women. *Diabet Med* 2011;28:23-30.
47 Joseph J, Svartberg J, Njolstad I, Schirmer H. Incidence of and risk factors for type-2 diabetes in a general population: the Tromso Study. *Scand J Public Health* 2010;38:768-75.
48 Von Eckardstein A, Schulte H, Assmann G. Risk for diabetes mellitus in middle-aged Caucasian male participants of the PROCAM study: implications for the definition of impaired fasting glucose by the American Diabetes Association. Prospective Cardiovascular Munster. *J Clin Endocrinol Metab* 2000;85:3101-8.
49 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;286:180-7.
50 Van Loon AJ, Tijhuis M, Picavet HS, Surtees PG, Ormel J. Survey non-response in the Netherlands: effects on prevalence estimates and associations. *Ann Epidemiol* 2003;13:105-10.
51 Herder C, Baumert J, Zierer A, Roden M, Meisinger C, Karakas M, et al. Immunological and cardiometabolic risk factors in the prediction of type 2 diabetes and coronary events: MONICA/KORA Augsburg case-cohort study. *PLoS One* 2011;6:e19852.
52 Van Dieren S, Nothlings U, van der Schouw YT, Spijkerman AM, Rutten GE, van der AD, et al. Non-fasting lipids and risk of cardiovascular disease in patients with diabetes mellitus. *Diabetologia* 2011;54:73-7.
53 Langenberg C, Sharp S, Forouhi NG, Franks PW, Schulze MB, Kerrison N, et al. Design and cohort description of the InterAct Project: an examination of the interaction of genetic and lifestyle factors on the incidence of type 2 diabetes in the EPIC Study. *Diabetologia* 2011;54:2272-82.
54 Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984;3:143-52.
55 McGeechan K, Macaskill P, Irwig L, Liew G, Wong TY. Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. *Arch Intern Med* 2008;168:2304-10.
56 Rathmann W, Martin S, Haastert B, Icks A, Holle R, Lowel H, et al. Performance of screening questionnaires and risk scores for undiagnosed diabetes: the KORA Survey 2000. *Arch Intern Med* 2005;165:436-41.
57 Schmidt MI, Duncan BB, Bang H, Pankow JS, Ballantyne CM, Golden SH, et al. Identifying individuals at high risk for diabetes: the Atherosclerosis Risk in Communities study. *Diabetes Care* 2005;28:2013-8.
58 Griffin SJ, Little PS, Hales CN, Kinmonth AL, Wareham NJ. Diabetes risk score: towards earlier detection of type 2 diabetes in general practice. *Diabetes Metab Res Rev* 2000;16:164-71.
59 Rahman M, Simmons RK, Harding AH, Wareham NJ, Griffin SJ. A simple risk score identifies individuals at high risk of developing type 2 diabetes: a prospective cohort study. *Fam Pract* 2008;25:191-6.

Accepted: 27 August 2012

Cite this as: *BMJ* 2012;345:e5900

Tables

Table 1| General characteristics of models to predict risk of incident type 2 diabetes included in study

Study and prediction risk model	Cases/total sample size	Ascertainment of incident diabetes*	Prediction horizon (years)	Statistical model	No of predictors	Discrimination C statistic (95% CI)	Calibration P value†
Alssema, 2011, Netherlands[25]							
DETECT-2	844/18 301	Self report, 2 h plasma glucose	5	Logistic	8	0.764 (0.746 to 0.783)	0.27
Wannamethee, 2011, UK[46]							
BRHS:							
Simple clinical	297/6927	Self report, review of patients' notes	7	Logistic	8	0.765 (0.740 to 0.791)	0.006
Fasting bio					11	0.817 (0.793 to 0.840)	0.43
Non-fasting bio					11	0.809 (0.785 to 0.833)	0.61
Rathmann, 2010, Germany[35]							
KORA:							
Basic	91/873	Self report, fasting glucose, non-fasting glucose	7.5	Logistic	6	0.763 (0.713 to 0.812)	0.66
Clinical					9	0.844 (0.801 to 0.887)	0.45
Chen, 2010, Australia[36]							
AUSDRISK	362/ 6060	Drug use, fasting glucose, 2 h plasma glucose	5	Logistic	9	0.78 (0.76 to 0.81)	0.85
Rosella, 2010, Canada[37]							
DPoRT	1410/19 861	Physician diagnosed in survey data	9	Weibull	7	M: 0.77 (0.76 to 0.79); F: 0.78 (0.76 to 0.79)	<0.01
Joseph, 2010, Norway[47]							
Tromsø	522/26 168	Self report, HbA_{1c}, medical record, fasting glucose	10.8	Cox	10	M: 0.87; F: 0.88	NR
Kahn, 2009, US[26]							
ARIC:							
Basic	1821/ 9587	Self report, fasting glucose, non-fasting glucose, hospital records, questionnaire	15	Weibull	10	0.71 (0.69 to 0.73)	NR
Enhanced					13	0.79 (0.77 to 0.81)	
Hippisley-Cox, 2009, UK[38]							
QDScore	78 081/2 540 753	General practice computer records	10	Cox	9	M: 0.834 (0.831 to 0.836); F: 0.853 (0.850 to 0.856)	Almost perfect calibration reported
Balkau, 2008, France[39]							
DESIR:							
Clinical	203/ 3814	Drug use, fasting glucose	9	Logistic	4	M: 0.733; F: 0.839	M: 0.7; F 0.6
Clinical+bio					6	M: 0.850; F: 0.917	M: 0.8; F: 0.9
Wilson, 2007, US[42]							
Framingham:							
Model 1	160/3140	Drug use, fasting glucose	7	Logistic	8	0.852	NR
Model 2					8	0.850	
Model 3					9	0.852	
Continuous					6	0.881	
Simmons, 2007, UK[41]							
EPIC-Norfolk	209/12 310	Hospital and general practice resisters, drug use, HbA_{1c} >7%	5	Logistic	9	0.762 (0.730 to 0.790)	NR
Schulze, 2007, Germany[40]							
GDRS (EPIC-Potsdam)	849/25 167	Self report verified by diagnosing physician	5	Cox	11	0.82 to 0.84	NR

Table 1 (continued)

Study and prediction risk model	Cases/total sample size	Ascertainment of incident diabetes*	Prediction horizon (years)	Statistical model	No of predictors	Discrimination C statistic (95% CI)	Calibration P value†
Lindstrom, 2003, Finland[43]							
FINDRISC:							
Concise	182/4435	Fasting glucose, non-fasting glucose	10	Logistic	5	0.857	NR
Full					7	0.860	
Stern, 2002, US[44]							
San Antonio, clinical	275/3004	Drug use, fasting glucose, 2 h plasma glucose	7.5	Logistic	8	0.843 (0.818 to 0.867)	>0.20
Von Eckardstein, 2000, Germany[48]							
PROCAM	200/3737	Self report, fasting glucose	6.3	Logistic	8	0.793 (0.780 to 0.806)	NR
Stern, 1993, US[45]							
San Antonio-reduced model	79/1453	Drug use, fasting glucose, 2 h plasma glucose	8	Logistic	5	NR	Excellent calibration reported

NR=not reported. DETECT-2=Detection Strategies for Type 2 Diabetes and Impaired Glucose Tolerance, KORA=Cooperative Health Research in Region of Augsburg; BRHS=British Regional Heart Study; AUDSRISK=Australian Type 2 Diabetes Risk Assessment Tool; DPoRT=Diabetes Population Risk Tool; ARIC=Atherosclerosis Risk in Communities; QDScore=diabetes risk algorithm; DESIR=Data from Epidemiological Study on Insulin Resistance Syndrome; EPIC=European Prospective Investigation Into Cancer and Nutrition; GDRS=German Diabetes Risk Score; FINDRISC=Finnish diabetes risk score; PROCAM=Prospective Cardiovascular Münster Study.

*Criteria for plasma glucose concentrations were as fasting glucose (7.0 mmol/L (126 mg/dL) and non-fasting or two hour glucose ≥11.1 mmol/L (200 mg/dL).

†Hosmer-Lemeshow χ^2.

Table 2| Discrimination and calibration of 12 basic models for prediction of risk of incident type 2 diabetes in validation cohort*

Risk prediction model	C statistic (95% CI)			Hosmer-Lemeshow χ^{2}*			O/E ratio at 7.5 years (95% CI)	Calibration slope†	Recalibrated Hosmer-Lemeshow χ^{2} at 7.5 years (P value)
	At 5 years	At 7.5 years	At 10 years	At 5 years	At 7.5 years	At 10 years			
DETECT-2, 2011	0.83 (0.82 to 0.85)	0.82 (0.81 to 0.84)	0.82 (0.81 to 0.83)	1704.8	1395.2	1090.0	0.304 (0.281 to 0.327)	0.87	275.5 (<0.001)
KORA, 2010, basic	0.83 (0.82 to 0.85)	0.83 (0.81 to 0.84)	0.82 (0.81 to 0.83)	1500.31	669.0	305.3	0.505 (0.468 to 0.545)	0.98	5.0 (0.17)
BRHS, 2011. simple clinical	0.80 (0.78 to 0.82)	0.79 (0.78 to 0.81)	0.79 (0.78 to 0.80)	532.9	685.7	841.1	0.740 (0.685 to 0.779)	0.26	356.5 (<0.001)
AUSDRISK, 2010	0.84 (0.83 to 0.86)	0.84 (0.82 to 0.85)	0.83 (0.82 to 0.84)	11 360.85	7195.7	4717.24	0.313 (0.290 to 0.337)	0.98	41.5 (<0.001)
DPoRT, 2010	0.75 (0.73 to 0.76)	0.74 (0.73 to 0.75)	0.74 (0.73 to 0.75)	245 226.8	173 309.8	120 485.1	0.060 (0.056 to 0.065)	—	864.8 (<0.001)
ARIC, 2009, basic	0.83 (0.82 to 0.85)	0.83 (0.81 to 0.84)	0.82 (0.81 to 0.84)	74 596.1	50 473.9	31 173.8	0.137 (0.127 to 0.147)	—	383.4 (<0.001)
QDScore, 2009	0.77 (0.75 to 0.79)	0.76 (0.74 to 0.78)	0.74 (0.72 to 0.76)	2176.5	1334.6	703.9	0.370 (0.365 to 0.375)	—	27.8 (<0.001)
DESIR, 2008, clinical	0.82 (0.80 to 0.84)	0.81 (0.80 to 0.83)	0.81 (0.79 to 0.82)	10 511.8	6342.3	3567.9	0.250 (0.232 to 0.270)	0.82	27.8 (<0.001)
EPIC-Norfolk , 2007	0.82 (0.80 to 0.84)	0.81 (0.80 to 0.82)	0.81 (0.79 to 0.82)	219.0	398.0	665.8	3.725 (3.450 to 4.016)	1.05	62.5 (<0.001)
EPIC-Potsdam, 2007, GDRS	0.84 (0.82 to 0.85)	0.84 (0.82 to 0.85)	0.83 (0.82 to 0.84)	569.2	152.7	49.9	0.805 (0.746 to 0.868)	—	67.8 (<0.001)
FINDRISC, 2003:									
Concise	0.83 (0.82 to 0.85)	0.82 (0.80 to 0.83)	0.81 (0.80 to 0.82)	1206.0	256.7	60.8	0.805 (0.746 to 0.868)	1.14	75.4 (<0.001)
Full	0.83 (0.81 to 0.85)	0.82 (0.80 to 0.83)	0.81 (0.80 to 0.82)	762.3	178.6	79.0	0.714 (0.661 to 0.769)	—	91.1 (<0.001)

O/E=observed to expected. DETECT-2=Detection Strategies for Type 2 Diabetes and Impaired Glucose Tolerance; KORA=Cooperative Health Research in Region of Augsburg; BRHS=British Regional Heart Study; AUDSRISK=Australian Type 2 Diabetes Risk Assessment Tool; DPoRT=Diabetes Population Risk Tool; ARIC=Atherosclerosis Risk in Communities; QDScore=diabetes risk algorithm; DESIR=Data from Epidemiological Study on Insulin Resistance Syndrome; EPIC=European Prospective Investigation Into Cancer and Nutrition; GDRS=German Diabetes Risk Score, FINDRISC Finnish diabetes risk score.

*All P<0.001.

†After recalibration.

Table 3| Discrimination and calibration of 13 extended models for prediction of risk of incident type 2 diabetes in validation cohort*

Risk prediction model	C statistic (95% CI)			Hosmer-Lemeshow χ^2†			O/E at 7.5 years (95% CI)	Calibration slope‡	Recalibrated Hosmer-Lemeshow χ^2 at 7.5 years†
	At 5 years	At 7.5 years	At 10 years	At 5 years	At 7.5 years	At 10 years			
BRHS, 2011:									
Fasting bio	0.87 (0.85 to 0.88)	0.86 (0.84 to 0.87)	0.86 (0.84 to 0.87)	951.4	1203.0	1530.8	0.857 (0.794 to 0.924)	0.44	115.9
Non-fasting bio	0.82 (0.81 to 0.84)	0.81 (0.80 to 0.83)	0.81 (0.80 to 0.82)	724.3	923.0	1153.4	0.361 (0.334 to 0.389)	0.44	116.3
KORA, 2010, clinical	0.94 (0.93 to 0.95)	0.93 (0.92 to 0.94)	0.92 (0.91 to 0.93)	1295.9	598.1	207.8	0.547 (0.507 to 0.590)	0.41	35.7
Tromsø, 2010	0.82 (0.81 to 0.84)	0.81 (0.80 to 0.83)	0.81 (0.80 to 0.82)	135 428.6	91 624.1	63 102.9	0.060 (0.055 to 0.064)	—	347.1
ARIC, 2009, enhanced	0.90 (0.88 to 0.91)	0.89 (0.87 to 0.90)	0.88 (0.87 to 0.89)	98 023.2	47 966.4	38 280.5	0.171 (0.159 to 0.185)	—	793.1
DESIR , 2008, clinical+bio	0.89 (0.87 to 0.90)	0.88 (0.87 to 0.89)	0.88 (0.87 to 0.89)	10 396.5	9565.1	8659.8	0.165 (0.153 to 0.178)	0.21	5649.5
Framingham, 2007:									
Model 1	0.82 (0.81 to 0.84)	0.82 (0.80 to 0.83)	0.81 (0.80 to 0.83)	2379.7	1300.3	490.7	0.487 (0.451 to 0.525)	—	125.1
Model 2	0.82 (0.80 to 0.84)	0.81 (0.80 to 0.83)	0.81 (0.80 to 0.83)	2818.6	1704.0	573.3	0.476 (0.441 to 0.513)	—	191.0
Model 3	0.83 (0.81 to 0.84)	0.82 (0.81 to 0.83)	0.82 (0.81 to 0.83)	3948.2	2503.3	904.3	0.423 (0.391 to 0.456)	—	203.7
Continuous	0.89 (0.88 to 0.90)	0.88 (0.87 to 0.89)	0.88 (0.86 to 0.89)	42 639.4	96407.4	65761.1	0.086 (0.079 to 0.092)	0.26	8686.72
San Antonio, 2002	0.92 (0.91 to 0.93)	0.91 (0.90 to 0.92)	0.90 (0.89 to 0.91)	108 660.9	56 994.3	37 528.8	0.113 (0.104 to 0.122)	0.41	33.2
PROCAM, 2000	0.85 (0.83 to 0.86)	0.84 (0.83 to 0.85)	0.83 (0.82 to 0.84)	831.7	1060	1272.5	22 600 (20 190 to 23 505)	1.72	223.4
San Antonio, 1993	0.90 (0.89 to 0.92)	0.89 (0.88 to 0.90)	0.88 (0.87 to 0.90)	21 390.0	14 663.6	9342.2	0.188 (0.174 to 0.202)	0.28	235.8

O/E=observed to expected; BRHS=British Regional Heart Study; KORA=Cooperative Health Research in Region of Augsburg; ARIC=Atherosclerosis Risk in Communities; DESIR=Data from Epidemiological Study on Insulin Resistance Syndrome; PROCAM=Prospective Cardiovascular Münster Study.

*Extrapolation approach applied for case cohort design.

†All P<0.001.

‡After recalibration.

Table 4| Performance of 12 basic models for prediction of risk of incident type 2 diabetes in sensitivity analyses

Risk prediction model	C statistic (95% CI) at 7.5 years			C statistic (95% CI) for original prediction horizon
	Case cohort design in extrapolated dataset*	Random glucose <11.1 mmol/L	Verified diabetes status	
DETECT-2, 2011	0.82 (0.81 to 0.84)	0.82 (0.81 to 0.84)	0.83 (0.81 to 0.84)	—
KORA, 2010, basic	0.82 (0.81 to 0.84)	0.82 (0.81 to 0.84)	0.83 (0.81 to 0.84)	—
BRHS, 2011, simple clinical	0.79 (0.77 to 0.80)	0.79 (0.77 to 0.80)	0.79 (0.78 to 0.81)	0.79 (0.78 to 0.81)†
AUSDRISK, 2010	0.84 (0.82 to 0.85)	0.83 (0.81 to 0.84)	0.83 (0.82 to 0.85)	—
DPoRT, 2010	0.74 (0.72 to 0.75)	0.74 (0.72 to 0.75)	0.74 (0.72 to 0.76)	0.74 (0.73 to 0.75)‡
ARIC, 2009, basic	0.82 (0.81 to 0.84)	0.83 (0.81 to 0.84)	0.83 (0.82 to 0.84)	0.82 (0.81 to 0.84)§
QDScore, 2009	0.76 (0.74 to 0.78)	0.76 (0.74 to 0.79)	0.76 (0.74 to 0.78)	—
DESIR, 2008, clinical	0.80 (0.79 to 0.82)	0.81 (0.79 to 0.82)	0.81 (0.80 to 0.83)	0.81 (0.79 to 0.82)‡
EPIC-Norfolk, 2007	0.81 (0.79 to 0.82)	0.81 (0.79 to 0.82)	0.81 (0.80 to 0.83)	—
EPIC-Potsdam, 2007, GDRS	0.83 (0.82 to 0.85)	0.83 (0.82 to 0.85)	0.84 (0.82 to 0.85)	—
FINDRISC, 2003:				
Concise	0.81 (0.80 to 0.83)	0.81 (0.80 to 0.83)	0.82 (0.80 to 0.83)	—
Full	0.81 (0.79 to 0.82)	0.81 (0.80 to 0.83)	0.82 (0.80 to 0.83)	—

DETECT-2=Detection Strategies for Type 2 Diabetes and Impaired Glucose Tolerance, KORA=Cooperative Health Research in Region of Augsburg; BRHS=British Regional Heart Study; AUDSRISK=Australian Type 2 Diabetes Risk Assessment Tool; DPoRT=Diabetes Population Risk Tool; ARIC=Atherosclerosis Risk in Communities; QDScore=diabetes risk algorithm; DESIR=Data from Epidemiological Study on Insulin Resistance Syndrome; EPIC=European Prospective Investigation Into Cancer and Nutrition; GDRS=German Diabetes Risk Score; FINDRISC=Finnish diabetes risk score.

*All Hosmer-Lemeshow χ^2 P<0.001.

†At 7 years.

‡At 9 years.

§At 15 years.

Table 5| **Performance of 13 extended models for prediction of risk of incident type 2 diabetes in sensitivity analyses***

Risk prediction model	C statistic (95% CI) at 7.5 years			C statistic (95% CI) for original prediction horizon
	Random glucose <11.1 mmol/L	MORGEN dataset	Fasting for >2 hours	
BRHS, 2011:				
Fasting bio	0.86 (0.84 to 0.88)	0.89 (0.87 to 0.91)	0.86 (0.84 to 0.88)	0.86 (0.84 to 0.87)†
Non-fasting bio	0.81 (0.80 to 0.83)	0.84 (0.82 to 0.86)	0.80 (0.79 to 0.82)	0.81 (0.80 to 0.83)†
KORA, 2010, clinical	0.93 (0.92 to 0.94)	0.92 (0.90 to 0.93)	0.93 (0.92 to 0.94)	—
Tromsø, 2010	0.81 (0.80 to 0.83)	0.84 (0.83 to 0.86)	0.81 (0.79 to 0.83)	0.81 (0.80 to 0.82)‡
ARIC, 2009 enhanced	0.88 (0.87 to 0.89)	0.91 (0.89 to 0.92)	0.90 (0.88 to 0.91)	0.88 (0.87 to 0.89)§
DESIR, 2008, clinical+bio	0.87 (0.86 to 0.88)	0.89 (0.87 to 0.91)	0.90 (0.88 to 0.91)	0.88 (0.87 to 0.89)¶
Framingham, 2007:				
Model 1	0.82 (0.80 to 0.83)	0.79 (0.77 to 0.81)	0.82 (0.80 to 0.84)	0.81 (0.80 to 0.83)†
Model 2	0.82 (0.80 to 0.83)	0.79 (0.76 to 0.81)	0.82 (0.80 to 0.84)	0.81 (0.80 to 0.83)†
Model 3	0.82 (0.81 to 0.84)	0.80 (0.77 to 0.82)	0.82 (0.80 to 0.84)	0.82 (0.81 to 0.83)†
Continuous	0.87 (0.86 to 0.88)	0.89 (0.87 to 0.91)	0.90 (0.88 to 0.91)	0.88 (0.87 to 0.89)†
San Antonio, 2002	0.90 (0.89 to 0.91)	0.91 (0.87 to 0.92)	0.92 (0.91 to 0.93)	—
PROCAM, 2000	0.84 (0.82 to 0.85)	0.87 (0.85 to 0.88)	0.83 (0.81 to 0.85)	0.84 (0.83 to 0.85)**
San Antonio, 1993	0.88 (0.87 to 0.90)	0.89 (0.86 to 0.91)	0.91 (0.90 to 0.92)	0.89 (0.87 to 0.90)††

BRHS=British Regional Heart Study; KORA=Cooperative Health Research in Region of Augsburg; ARIC=Atherosclerosis Risk in Communities; DESIR=Data from Epidemiological Study on Insulin Resistance Syndrome; PROCAM=Prospective Cardiovascular Münster Study.

*Extrapolation approach applied for case cohort design.

†At 7 years.

‡At 10.8 years.

§At 15 years.

¶At 9 years.

**At 6.3 years.

††At 8 years.

Figures

Fig 1 Overview of systematic literature search of studies that derived prediction models for risk of type 2 diabetes

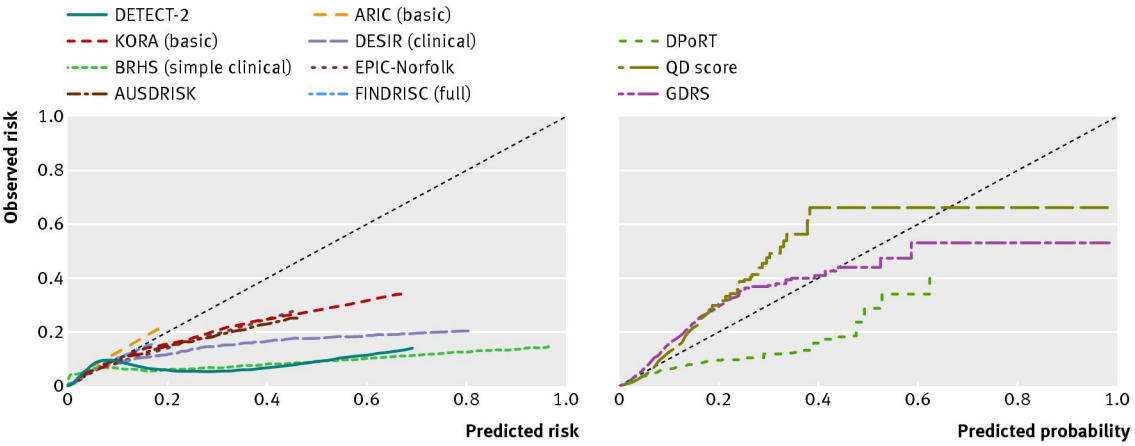

Fig 2 Calibration plots for recalibrated basic models risk of diabetes at 7.5 years depicting predicted risk against observed risk of developing type 2 diabetes in validation dataset. Dashed line (45° line) from zero denotes ideal calibration line (slope=1, intercept=0) and other lines are smooth calibration curve for each model

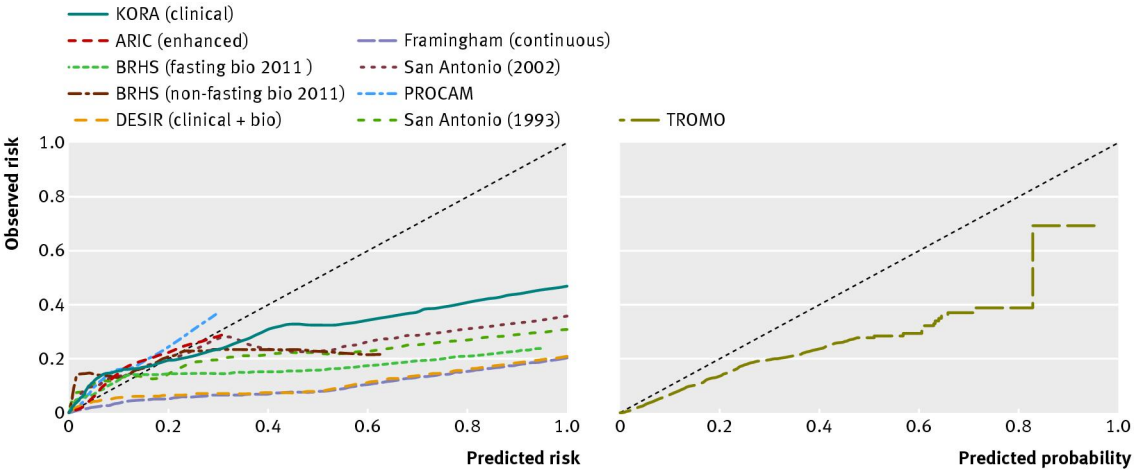

Fig 3 Calibration plots for recalibrated extended models risk of diabetes at 7.5 years depicting predicted risk against observed risk of developing type 2 diabetes in validation dataset. Dashed line (45° line) from zero denotes ideal calibration line (slope=1, intercept=0) and other lines are smooth calibration curve for each model