Kengne, AP; Beulens, JWJ; Peelen, LM; Moons, KGM; van der Schouw, YT; Schulze, MB; Spijkerman, AMW; Griffin, SJ; Grobbee, DE; Palla, L; Tormo, M.-, J; Arriola, L; Barengo, NC; Barricarte, A; Boeing, H; Bonet, C; Clavel-Chapelon, F; Dartois, L; Fagherazzi, G; Franks, PW; Huerta, JM; Kaaks, R; Key, TJ; Khaw, KT; LI, K; Muhlenbruch, K; Nilsson, PM; Overvad, K; Overvad, TF; Palli, D; Panico, S; Quiros, JR; Rolandsson, O; Roswall, N; Sacerdote, C; Sanchez, M.-, J; Slimani, N; Tagliaabue, G; Tjonneland, A; Tumino, R; van der A, DL; Forouhi, NG; Sharp, SJ; Langenberg, C; Riboli, E; Wareham, NJ (2013) Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. The lancet Diabetes & endocrinology, 2 (1). pp. 19-29. ISSN 2213-8587 DOI: https://doi.org/10.1016/S2213-8587(13)70103-7

Downloaded from: http://researchonline.lshtm.ac.uk/1924654/

DOI: 10.1016/S2213-8587(13)70103-7

Usage Guidelines

Please refer to usage guidelines at http://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by/2.5/
Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models


Summary

Background The comparative performance of existing models for prediction of type 2 diabetes across populations has not been investigated. We validated existing non-laboratory-based models and assessed variability in predictive performance in European populations.

Methods We selected non-invasive prediction models for incident diabetes developed in populations of European ancestry and validated them using data from the EPIC-InterAct case-cohort sample (27 779 individuals from eight European countries, of whom 12 403 had incident diabetes). We assessed model discrimination and calibration for the first 10 years of follow-up. The models were first adjusted to the country-specific diabetes incidence. We did the main analyses for each country and for subgroups defined by sex, age (<60 years vs ≥60 years), BMI (<25 kg/m² vs ≥25 kg/m²), and waist circumference (men <102 cm vs ≥102 cm; women <88 cm vs ≥88 cm).

Findings We validated 12 prediction models. Discrimination was acceptable to good: C statistics ranged from 0·76 (95% CI 0·72–0·80) to 0·81 (0·77–0·84) overall, from 0·73 (0·70–0·76) to 0·79 (0·74–0·83) in men, and from 0·78 (0·74–0·82) to 0·81 (0·80–0·82) in women. We noted significant heterogeneity in discrimination (p heterogeneity <0·0001) in all but one model. Calibration was good for most models, and consistent across countries (p heterogeneity >0·05) except for three models. However, two models overestimated risk, DProRT by 34% (95% CI 29–39%) and Cambridge by 40% (28–52%). Discrimination was always better in individuals younger than 60 years or with a low waist circumference than in those aged at least 60 years or with a large waist circumference. Patterns were inconsistent for BMI. All models overestimated risks for individuals with a BMI of <25 kg/m². Calibration patterns were inconsistent for age and waist-circumference subgroups.

Interpretation Existing diabetes prediction models can be used to identify individuals at high risk of type 2 diabetes in the general population. However, the performance of each model varies with country, age, sex, and adiposity.

Funding The European Union.

Introduction

The number of individuals with type 2 diabetes is high and increasing rapidly: 366 million people worldwide were estimated to have type 2 diabetes in 2011, and prevalence is expected to rise by 51% by 2030.1 Diabetes is associated with increased morbidity and mortality, and accounts for a substantial proportion of use of health-care resources worldwide.2 Several studies have convincingly shown that early interventions can prevent or postpone type 2 diabetes.3 The cost of these interventions and their constraints for individuals vary in many ways, such as the time horizon for prediction and number and nature of predictors. Some models are based on non-laboratory clinical variables (non-invasive risk scores); others have also incorporated biological variables (invasive risk scores). Non-invasive risk scores have been shown to identify a high risk of type 2 diabetes (C statistics ≥0·8), although invasive risk scores are more successful (C statistics ≥0·9).4 The use of non-invasive risk scores is more likely to be cost effective and feasible for large-scale screening than is use of invasive risk scores. Generally, the use of risk scores has been widely incorporated into strategies for diabetes prevention.5,6 However, risk scores might not be generalisable from one population to another, and the validity of existing diabetes risk scores in different settings and subgroups on the basis of sex or BMI has not been established.7

We aimed to validate and compare existing non-invasive prediction models for type 2 diabetes in European populations. We assessed variability in predictive performance between countries and by sex, BMI, waist circumference, and age.

Methods
Study design and population
We used the EPIC-InterAct case-cohort sample (27,779 participants, of whom 12,403 had incident diabetes), which was selected from the European Investigation into Cancer and Nutrition (EPIC) cohort study (455,569 participants; appendix p 4).15 The design and methods of the InterAct study have been described elsewhere.16 Briefly, the InterAct study was designed to investigate how genes and lifestyle factors interact to influence risk of type 2 diabetes. The InterAct consortium partners ascertained and verified incident cases of type 2 diabetes in EPIC cohorts between 1991 and 2007, from eight of the ten EPIC countries (26 centres).

Participants gave written informed consent for participation in the EPIC study. The study was approved by the local ethics committees in the participating countries and the internal review board of the International Agency for Research on Cancer (Lyon, France).

Procedures
Self-administered questionnaires provided baseline information about lifestyle and medical history in the case-cohort sample. Validated instruments were used to assess nutritional variables12 and physical activity.13 Blood pressure and anthropometric measurements followed standard approaches (appendix p 5). We handled missing data in three ways: proxy variables (appendix p 6) or exclusion of the predictor from the model; exclusion of countries from specific analyses (appendix p 11); and imputation of missing values by country with R’s AregImpute Function, which takes all aspects of uncertainty in the imputations into account by using the bootstrap to estimate the process of drawing predicted values from a full Bayesian predictive distribution.

Existing non-invasive prediction models for incident diabetes have been identified through a systematic review, as previously described.8 We included in this study only models developed in general populations of European ancestry. We first validated the original models by computing the predicted probability of diabetes for each participant in the EPIC-InterAct case-cohort with baseline values (appendix p 7). We expressed models’ performance in terms of discrimination (whether the model can distinguish between people who do and do not develop diabetes) and calibration (to what extent the predicted probabilities agree with the reported risk across groups of individuals). We assessed discrimination with the C statistic (which is comparable to the area under the receiver operator characteristic curve) adapted for time-to-event data.17 C statistics vary from 0·5 (no discrimination) to 1 (perfect discrimination), with values of 0·7–0·8 deemed acceptable and 0·8–0·9 good.18 We assessed calibration graphically with calibration plots, and by computing the ratio of expected to recorded probabilities and the accompanying 95% CIs by assuming a Poisson variance.19 We also calculated the Yates slope (difference between mean predicted probability of type 2 diabetes for participants with and without incident diabetes, with higher values indicating better performance), and Brier score (squared difference between predicted probability and actual outcome for each participant; these values vary from 0 for a perfect prediction model to 0–25 for a non-informative model with 50% incident outcome).20 We did not use tests of calibration such as the Hosmer-Lemeshow test or U statistic,21 because they are sensitive to study sample size.

Because we were comparing the performance of several models in a case-cohort design across many countries, we took some extra steps before model validation. First, the models have different so-called time horizons. In the main analysis, we compared all models for prediction of type 2 diabetes within 10 years. Therefore, participants who developed diabetes after 10 years of follow-up were included as non-cases. Second, calibration is strongly affected by the outcome’s incidence in the population; we therefore recalibrated models to the country-specific incidence through intercept adjustment (appendix pp 9, 12). These approaches eliminate differences between the models due to differences in incidence between the populations for which the scores were developed.22 Third, the incidence of diabetes is artificially increased in the case-cohort design. To arrive at the true incidences from the original cohort, we reconstituted the EPIC cohort within countries by applying a so-called blow-up approach to extrapolate the case-cohort data to a full cohort (appendix p 10). Finally, to account for any heterogeneity between countries, we estimated the C statistic and ratio of expected to recorded probabilities for the entire cohort by pooling country-level estimates with random effects models. We then used the Cochran’s Q and I² statistics to assess heterogeneity across countries.

We did the main analyses for each country and for subgroups defined by sex, age (<60 years vs ≥60 years), BMI (<25 kg/m² vs ≥25 kg/m²), and waist circumference (men <102 cm vs ≥102 cm; women <88 cm vs ≥88 cm). In four of the eight countries (Denmark, Italy, Spain, and Sweden), not all models had been validated (appendix p 11). We did a sensitivity analysis by restricting the assessment of the heterogeneity between countries to the four countries in which all models had been validated (France, Germany, the Netherlands, and the UK) to investigate if assessments of heterogeneity of different models on the basis of a different number of countries affected the comparisons. We also compared all models for prediction of diabetes within 5 years, which corresponds to the time horizon for predicted probability for six models.

We used the Hmisc, Design, Survival, and meta packages of R (version 2.13.0) for data analysis.
Role of the funding source
The sponsors of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. APK and JWJB had full access to all the data in the study; JWJB, YTvdS, and NJW had final responsibility for the decision to submit for publication.

Results
We validated 12 prediction models (table 1, appendix p 7). Age was a predictor of type 2 diabetes in all models except DESIR.22 Most models were based on logistic regressions (table 1). Varying definitions for incident diabetes were applied (table 1). Data for baseline age and sex were available for all participants, but data were missing for some participants for some predictors (table 1).

More than half the overall EPIC-InterAct study population were women, but the proportion varied by country (table 2). French participants were all women, whereas the proportion of women among participants in other countries ranged from 44% to 81%. Mean baseline age was 54·1 years (SD 8·4) in men and 53·5 years (9·0) in women.

During the first 10 years of follow-up, almost 10 000 cases of incident diabetes were recorded in the EPIC-InterAct population (table 2). Discrimination of the 12 prediction models was acceptable to good, with C statistics ranging from 0·76 to 0·81 (figure 1, table 3). For nine models, the recalibrated calibration curves were mostly followed the ideal calibration line (figure 2; see appendix p 21 for calibration curves before recalibration). For the ARIC 2009 model,21 the curves were mostly steeper than the ideal calibration line, with increasing risk (figure 2), indicating risk underestimation in high-risk participants. For the Cambridge23 and DPoRT25 models, calibration curves were fairly flat, with increasing risk (figure 2), indicating risk overestimation in high-risk participants. Four models accurately estimated the overall rate of incident diabetes, six marginally overestimated or underestimated, and two largely overestimated (figure 2, table 3).

We noted significant heterogeneity in discrimination across countries for all models except for the ARIC 2009 model (figure 1). The highest C statistics were almost always recorded in France, and the lowest mostly in Denmark (figure 1). The worst and best discriminating models varied by country (figure 1). Patterns of calibration were largely similar in each country to those recorded for the overall cohort (appendix pp 22–33). As expected, because of recalibration by country, we recorded almost no heterogeneity in calibration.

Discrimination was generally higher in women than men (table 3). C statistics in men ranged from 0·73 to 0·79 in the overall cohort (table 3), and from 0·68 to 0·86 across countries (appendix p 13). In men, the best discrimination was recorded in the Netherlands (appendix p 13). In women, C statistics ranged from 0·78 to 0·81 in the overall cohort (table 3), and from 0·70 to 0·85 across countries (appendix p 13). The highest C statistics in women were mostly recorded in Germany, and the lowest in Denmark (appendix p 13). Calibration differed greatly between men and women per country, which could be masked by the overall ratio of expected to recorded probabilities (appendix p 13).

Discrimination was acceptable to good across age strata, and was always better in the younger age group.
than in the older (table 3). Discrimination was below the acceptable range (C statistic < 0.70) in the lower stratum of waist circumference but was acceptable in the upper one. Discrimination was similar across BMI groups, although C statistics were higher in people with increased BMI than in those with lower BMIs in seven of the 12 models (table 3).

In terms of calibration, the ARIC 2005, 20 AUSDRISK, 22 FINDRISK concise, and FINDRISK full models acceptably predicted the overall rate of incident diabetes in age subgroups (table 3). The Cambridge model 23 overestimated risk in all participants (table 3). Most other models overestimated risk in younger participants and underestimated risk in the older people, but we recorded...
the opposite finding for KORA and Potsdam models. Absolute risk of diabetes was variably overestimated or underestimated by models across weight circumference strata, while all models systematically overestimated the risk within the lower BMI stratum (table 3).

The pattern of overall event rate prediction in the four countries with valid data for all models (France, Germany, the Netherlands, and the UK) was very similar to that across all countries (data not shown), again with significant heterogeneity for the Cambridge and across all countries (data not shown), again with the Netherlands, and the UK) was very similar to that countries with valid data for all models (France, Germany, Italy, Netherlands, Spain, Sweden, and the UK).

Absolute risk of diabetes was more accurately estimated in men than in women. Discrimination is better in people younger than 60 years, but risk can be overestimated in this age group. Discrimination is generally lower in participants with a BMI of less than 25 kg/m² than in those with a BMI of at least 25 kg/m². Risk is systematically overestimated in participants with a BMI of less than 25 kg/m². No model significantly outperforms others enough to be uniquely recommended for routine risk stratification.

A few previous validation studies of incident diabetes models have been done. Most of those—mainly cross-sectional studies—compared several incident or prevalent risk scores in one population, and another study compared three risk scores in a multiethnic population in the same country. The previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.


calibration, which was consistent across countries, although discrimination varied significantly. We showed that the models' performance is worse in men than in women. Discrimination is better in people younger than 60 years, but risk can be overestimated in this age group. Discrimination is generally lower in participants with a BMI of less than 25 kg/m² than in those with a BMI of at least 25 kg/m². Risk is systematically overestimated in participants with a BMI of less than 25 kg/m². No model significantly outperforms others enough to be uniquely recommended for routine risk stratification.

A few previous validation studies of incident diabetes models have been done. Most of those—mainly cross-sectional studies—compared several incident or prevalent risk scores in one population, and another study compared three risk scores in a multiethnic population in the same country. The previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.

Models' performance (mostly discrimination) differed across countries in our study. This variation could be a result of differences in how predictors and outcomes were assessed. Overall, the previous studies showed that the performance of both invasive and non-invasive prediction models for incident diabetes in a population was assessed. Overall, the previous studies showed that the models had modest to good discrimination and poor calibration. However, intercept adjustment to correct for differences in diabetes incidence between development and validation populations was done in only one study.
Figure 1: Discrimination of the different models for the prediction of incident type 2 diabetes at 10 years of follow-up overall and by country

Discrimination refers to the ability of the model to distinguish between participants who did and did not develop diabetes during 10 years of follow-up. ARIC=Atherosclerosis Risk in Communities. AUSDRISK=Australian Type 2 Diabetes Risk Assessment Tool. DESIR=Epidemiological Study on the Insulin Resistance Syndrome. DPoRT=Diabetes Population Risk Tool. FINDRISK=FINnish Diabetes Risk Score. KORA S4/F4=Cooperative Health Research in the Region of Augsburg (KORA), Survey 4.
were measured (for discrimination) and in baseline diabetes risk (for calibration) across countries. The FINDRISK concise model was previously tested in five cohorts from Europe, North America, and Australia.\textsuperscript{16} Discrimination varied by country, with C statistics between 0.63 and 0.78,\textsuperscript{15} which is consistent with our

<table>
<thead>
<tr>
<th>Overall</th>
<th>Men</th>
<th>Women</th>
<th>Age &lt;60 years</th>
<th>Age ≥60 years</th>
<th>BMI &lt;25 kg/m²</th>
<th>BMI ≥25 kg/m²</th>
<th>Waist circumference &lt;102 cm in men</th>
<th>Waist circumference ≥102 cm in men</th>
<th>Waist circumference ≥88 cm in women</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARK 2005</td>
<td>0.79 (0.76-0.83)\textsuperscript{*}</td>
<td>0.75 (0.73-0.78)\textsuperscript{*}</td>
<td>0.81 (0.77-0.84)\textsuperscript{*}</td>
<td>0.81 (0.77-0.85)\textsuperscript{*}</td>
<td>0.73 (0.69-0.76)\textsuperscript{*}</td>
<td>0.71 (0.69-0.73)\textsuperscript{*}</td>
<td>0.66 (0.65-0.67)\textsuperscript{†}</td>
<td>0.75 (0.70-0.79)\textsuperscript{*}</td>
<td></td>
</tr>
<tr>
<td>ARK 2009</td>
<td>0.80</td>
<td>0.79</td>
<td>0.81 (0.80-0.82)</td>
<td>0.81 (0.81-0.83)</td>
<td>0.74 (0.71-0.77)</td>
<td>0.71 (0.71-0.74)</td>
<td>0.66 (0.65-0.69)</td>
<td>0.75 (0.70-0.76)</td>
<td></td>
</tr>
<tr>
<td>AUSDRISK</td>
<td>0.77</td>
<td>0.73</td>
<td>0.78</td>
<td>0.78</td>
<td>0.73 (0.68-0.77)</td>
<td>0.71 (0.68-0.73)</td>
<td>0.64 (0.60-0.69)</td>
<td>0.69 (0.65-0.73)</td>
<td></td>
</tr>
<tr>
<td>Cambridge model</td>
<td>0.79 (0.75-0.82)</td>
<td>0.76 (0.73-0.79)</td>
<td>0.80 (0.76-0.83)</td>
<td>0.80 (0.76-0.84)</td>
<td>0.72 (0.70-0.74)</td>
<td>0.70 (0.68-0.72)</td>
<td>0.66 (0.64-0.68)</td>
<td>0.74 (0.69-0.79)</td>
<td></td>
</tr>
<tr>
<td>DESIR model</td>
<td>0.79 (0.76-0.81)</td>
<td>0.73 (0.71-0.75)</td>
<td>0.81 (0.78-0.85)</td>
<td>0.81 (0.77-0.85)</td>
<td>0.74 (0.71-0.77)</td>
<td>0.71 (0.65-0.76)</td>
<td>0.69 (0.64-0.74)</td>
<td>0.72 (0.67-0.76)</td>
<td></td>
</tr>
<tr>
<td>DnR model</td>
<td>0.76</td>
<td>0.73</td>
<td>0.78</td>
<td>0.78</td>
<td>0.70</td>
<td>0.70 (0.68-0.73)</td>
<td>0.68 (0.66-0.69)</td>
<td>0.64 (0.60-0.72)</td>
<td></td>
</tr>
<tr>
<td>FINDRISK concise</td>
<td>0.78 (0.74-0.79)</td>
<td>0.75 (0.71-0.78)</td>
<td>0.79 (0.75-0.81)</td>
<td>0.79 (0.76-0.82)</td>
<td>0.71 (0.69-0.74)</td>
<td>0.70 (0.69-0.72)</td>
<td>0.69 (0.65-0.73)</td>
<td>0.72 (0.67-0.74)</td>
<td></td>
</tr>
<tr>
<td>FINDRISK full model</td>
<td>0.77</td>
<td>0.74 (0.72-0.77)</td>
<td>0.78 (0.76-0.81)</td>
<td>0.78 (0.75-0.81)</td>
<td>0.71 (0.69-0.74)</td>
<td>0.70 (0.69-0.72)</td>
<td>0.69 (0.65-0.72)</td>
<td>0.71 (0.67-0.74)</td>
<td></td>
</tr>
<tr>
<td>Framingham personal</td>
<td>0.76</td>
<td>0.74</td>
<td>0.78</td>
<td>0.78</td>
<td>0.69</td>
<td>0.69 (0.67-0.73)</td>
<td>0.65 (0.63-0.69)</td>
<td>0.61 (0.57-0.67)</td>
<td></td>
</tr>
<tr>
<td>KORA model</td>
<td>0.78 (0.74-0.82)</td>
<td>0.75 (0.71-0.79)</td>
<td>0.79 (0.75-0.82)</td>
<td>0.79 (0.76-0.84)</td>
<td>0.72 (0.69-0.74)</td>
<td>0.71 (0.68-0.75)</td>
<td>0.68 (0.65-0.71)</td>
<td>0.70 (0.65-0.72)</td>
<td></td>
</tr>
<tr>
<td>Potsdam model</td>
<td>0.79 (0.76-0.82)</td>
<td>0.74 (0.72-0.77)</td>
<td>0.81 (0.78-0.83)</td>
<td>0.80 (0.77-0.83)</td>
<td>0.72 (0.69-0.75)</td>
<td>0.70 (0.68-0.75)</td>
<td>0.68 (0.65-0.71)</td>
<td>0.70 (0.65-0.73)</td>
<td></td>
</tr>
<tr>
<td>QDScore</td>
<td>0.81</td>
<td>0.78 (0.74-0.82)</td>
<td>0.81 (0.77-0.85)</td>
<td>0.82 (0.78-0.86)</td>
<td>0.72 (0.70-0.75)</td>
<td>0.72 (0.69-0.75)</td>
<td>0.68 (0.66-0.70)</td>
<td>0.72 (0.68-0.72)</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{*}Articles
results. Therefore, risk scores might not always be generalisable to populations other than the development population.

We noted some consistent trends across countries. Discrimination was generally low for the DPoRT model, potentially because this model was developed to estimate diabetes risk in a population rather than for an individual, which could affect quality of the predictors and outcome data. The QDScore was one of the best discriminatory models across countries. It is based on more predictors than other models are and so uses more information, and was developed from a large sample, providing precise estimates of the associations between predictors and outcomes. Calibration of the Cambridge model was generally poor. Although validated for incident diabetes prediction, this model was developed as a diagnostic model, which could have inflated estimates of the association between diabetes and predictors.

We recorded substantial variation in models' performance by major subgroups, with consistently better performance in women than in men. For the DESIR and DPoRT models and the QDScore, this difference could be a result of the use of sex-specific coefficients that could efficiently capture the predictive information from risk factors. BMI and waist circumference are generally much stronger diabetes predictors in women than in men and could also explain these results. In our study, the differences in discrimination linked to BMI were small. BMI and waist circumference might not be linearly related to diabetes risk. By including linear or categorical terms of these predictors, most models might not completely capture the nature of the associations. Our finding that discrimination is mostly better in younger than in older participants is consistent with a previous validation of the KORA model. However, stratification for a variable with discriminatory value by itself reduces discrimination.

The European guideline for the prevention of type 2 diabetes and the International Diabetes Federation recommend the use of reliable, straightforward, and practical prediction models to identify people at high risk of diabetes. Overall, our study shows that prediction models for type 2 diabetes are valid instruments for the identification of individuals at high risk. However, this finding cannot directly be translated into a recommendation to use a specific model, because the decision to adopt a prediction model in any specific setting is driven by many factors, of which performance is only one.

For individuals, highly accurate risk scores including biochemical testing are probably more relevant than non-invasive risk models are. However, non-invasive risk scores could be used as part of the public health approach to diabetes prevention to identify individuals who should receive biochemical testing. Alternatively, a risk score could be used to identify a subgroup of the population for lifestyle intervention without additional testing. Mortality risk is increased in the large group of people who have positive risk scores, justifying direct action in this group. Risk scores like those in the Cambridge model and QDScore in the UK, or the DPoRT model in Canada, use data available from records in the countries where models were developed, and new data are not needed. Because such scores are likely to be applicable to anyone coming into contact with the health system, their performance could be better overall than that of another model with a slightly better discrimination. Nevertheless, a model for which a questionnaire is designed for self-screening at the population level (eg, the two FINDRISK models) might not be based on many responses, potentially overriding any performance advantage recorded in our analysis.

Therefore, the decision to use a particular model could be country specific and depends on factors other than model performance, such as availability of measurements in the setting where the model is used. Additionally, prediction of diabetes is useful only when used to implement interventions to affect the outcomes of high-risk people. Diabetes prevention trials have been mostly based on a high-risk status defined by blood tests. However, implementation studies are necessary to investigate effectiveness of such risk scores. A few implementation studies have used risk scores to select participants for interventions and have shown favourable effects on risk factor levels—eg, a moderate weight loss effectively reduced risk of type 2 diabetes after 1 year of intervention in participants selected with the FINDRISK score.

Our study has some limitations. Some of the approaches used to account for predictors that were completely missing (ie, proxy variables, predictor omission, and country's exclusion) could marginally decrease discrimination of the models. It is now widely accepted that deletion of participants with missing values (who are frequent in large studies) yields biased results. Therefore, we applied robust methods to deal with missing data. The large sample and the multicountry nature of our study has allowed us to assess and compare 12 models across countries for the first time (panel). However, we did not include non-European participants. Therefore, our results might not be generalisable to non-European countries or other

Figure 2: Calibration curves for each model for the prediction of incident type 2 diabetes at 10 years of follow-up for the total cohort

Calibration of a model describes the extent to which the expected probability of diabetes matches the recorded probability of diabetes during follow-up. The ideal calibration (perfect agreement) is graphically represented by the dotted diagonal line at 45°. The vertical lines at the bottom of the graph depict the frequency distribution of the calibrated probabilities of diabetes. Grouped observations are for groups of participants across increasing deciles of predicted risk. Atherosclerosis Risk in Communities. AUSDRISK=Australian Type 2 Diabetes Risk Assessment Tool. DESIR=Epidemiological Study on the Insulin Resistance Syndrome. DPoRT=Diabetes Population Risk Tool. FINDRISK=Finnish Diabetes Risk Score. KORA S4/F4=Cooperative Health Research in the Region of Augsburg (KORA), Survey 4.
Articles

Panel: Research in context

Systematic review

We searched PubMed for studies published in English before July 31, 2013, with the search terms “diabetes”, “risk score”, “prediction model”, “prediction”, “risk assessment”, “algorithm”, “validation”, “calibration”, and “discrimination”. Previous validation studies of diabetes prediction models have tested the performance of one model in one population or compared the performance of a few models within the same population or country,4,5,12,14,30–34 but none has assessed and compared the performance of several models within and across several countries.

Interpretation

Our findings support existing recommendations that risk scores for type 2 diabetes can be used to identify individuals at high risk and start interventions to lower their future risk of type 2 diabetes. Available non-laboratory-based diabetes risk scores have acceptable to good discriminatory power in the ranking of individuals on the basis of risk across diverse settings in Europe. However, model performance differed across countries and no model outperformed the others enough to be uniquely recommended. Our results provide information to identify an appropriate risk score for each country.

Contributors

All authors contributed to study conception and design, data interpretation, and critical revision of the report. APK, JWJB, and LMP analysed the data and wrote the first draft of the report.

Conflicts of interest

SG received an honorarium and reimbursement of travel expenses from Eli Lilly, associated with membership of an independent data monitoring committee for a randomised trial of a drug to lower blood glucose concentrations. All other authors declare that they have no conflicts of interest.

Acknowledgments

The InterAct study received funding from the European Union (Integrated Project LSHTM-CT-2006-037197 in the Framework Programme 6 of the European Community). JWJB and YVdS received additional funding for verification of diabetes cases from the NL Agency (grant IGE95010) and an Incentive Grant from the Board of the University Medical Center Utrecht. AMWS and DivDA received funding from the Dutch Ministry of Public Health, Welfare and Sports; Netherlands Cancer Registry; LR Research Funds; Dutch Prevention Funds; Zorg Onderzoek Nederland; World Cancer Research Fund; and Statistics Netherlands. MJT received funding from the Health Research Fund of the Spanish Ministry of Health and the CIBER en Epidemiología y Salud Pública, and Murcia Regional Government (number 6236). PWF and FN received funding from the Swedish Research Council. PWF also received funding from Novo Nordisk, Swedish Diabetes Association, and Swedish Heart-Lung Foundation. RK and KL received funding from German Cancer Aid and the Federal Ministry of Education and Research. TK and KTK received funding from Cancer Research UK. KTK also received funding from Medical Research Council UK. KO and AT received funding from the Danish Cancer Society, OR received funding from the county of Västerbotten. JRQ and received funding from Asturias Regional Government. RT received funding from AIRE-ONLUS Ragusa, AVIS-Ragusa, and the Sicilian Regional Government. We thank all EPIC participants and staff for their contribution to the study.

References
