BMJ

RESEARCH

Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore

Julia Hippisley-Cox, professor of clinical epidemiology and general practice,[1] Carol Coupland, senior lecturer in medical statistics,[1] John Robson, senior lecturer in general practice,[2] Aziz Sheikh, professor of primary care research and development,[3] Peter Brindle, research and development strategy lead[4]

[1]Division of Primary Care, Tower Building, University Park, Nottingham NG2 7RD

[2]Centre for Health Sciences, Queen Mary's School of Medicine and Dentistry, London E1 2AT

[3]Centre for Population Health Sciences: GP Section, University of Edinburgh, Edinburgh EH8 9DX

[4]Avon Primary Care Research Collaborative, Bristol Primary Care Trust, Bristol BS2 8EE

Correspondence to: J Hippisley-Cox
Julia.hippisley-cox@nottingham.ac.uk

Cite this as: *BMJ* 2009;338:b880
doi:10.1136/bmj.b880

ABSTRACT

Objective To develop and validate a new diabetes risk algorithm (the QDScore) for estimating 10 year risk of acquiring diagnosed type 2 diabetes over a 10 year time period in an ethnically and socioeconomically diverse population.

Design Prospective open cohort study using routinely collected data from 355 general practices in England and Wales to develop the score and from 176 separate practices to validate the score.

Participants 2 540 753 patients aged 25-79 in the derivation cohort, who contributed 16 436 135 person years of observation and of whom 78 081 had an incident diagnosis of type 2 diabetes; 1 232 832 patients (7 643 037 person years) in the validation cohort, with 37 535 incident cases of type 2 diabetes.

Outcome measures A Cox proportional hazards model was used to estimate effects of risk factors in the derivation cohort and to derive a risk equation in men and women. The predictive variables examined and included in the final model were self assigned ethnicity, age, sex, body mass index, smoking status, family history of diabetes, Townsend deprivation score, treated hypertension, cardiovascular disease, and current use of corticosteroids; the outcome of interest was incident diabetes recorded in general practice records. Measures of calibration and discrimination were calculated in the validation cohort.

Results A fourfold to fivefold variation in risk of type 2 diabetes existed between different ethnic groups. Compared with the white reference group, the adjusted hazard ratio was 4.07 (95% confidence interval 3.24 to 5.11) for Bangladeshi women, 4.53 (3.67 to 5.59) for Bangladeshi men, 2.15 (1.84 to 2.52) for Pakistani women, and 2.54 (2.20 to 2.93) for Pakistani men. Pakistani and Bangladeshi men had significantly higher hazard ratios than Indian men. Black African men and Chinese women had an increased risk compared with the corresponding white reference group. In the validation dataset, the model explained 51.53% (95% confidence interval 50.90 to 52.16) of the variation in women and 48.16% (47.52 to 48.80) of that in men. The risk score showed good discrimination, with a D statistic of 2.11 (95% confidence interval 2.08 to 2.14) in women and 1.97 (1.95 to 2.00) in men. The model was well calibrated.

Conclusions The QDScore is the first risk prediction algorithm to estimate the 10 year risk of diabetes on the basis of a prospective cohort study and including both social deprivation and ethnicity. The algorithm does not need laboratory tests and can be used in clinical settings and also by the public through a simple web calculator (www.qdscore.org).

INTRODUCTION

The prevalence of type 2 diabetes and the burden of disease caused by it have increased very rapidly worldwide.[1] This has been fuelled by ageing populations,[2] poor diet,[3] and the concurrent epidemic of obesity.[4 5] The health and economic consequences of this diabetes epidemic are huge and rising.[6] Strong evidence from randomised controlled trials shows that behavioural or pharmacological interventions can prevent type 2 diabetes in up to two thirds of high risk cases.[7-10] Cost effectiveness modelling suggests that screening programmes aid earlier diagnosis and help to prevent type 2 diabetes or improve outcomes in people who develop the condition,[11 12] making the prevention and early detection of diabetes an international public health priority.[13 14] Early detection is important, as up to half of people with newly diagnosed type 2 diabetes have one or more complications at the time of diagnosis.[15]

Although several algorithms for predicting the risk of type 2 diabetes have been developed,[16-19] no widely accepted diabetes risk prediction score has been developed and validated for use in routine clinical practice. Previous studies have been limited by size,[16] and some have performed inadequately when tested in ethnically diverse populations.[20] A new diabetes risk prediction tool with appropriate weightings for both social deprivation and ethnicity is needed given the prevalence of type 2 diabetes, particularly among minority ethnic communities, appreciable numbers of whom remain without a diagnosis for long periods of time.[21] Such patients have an increased risk of avoidable morbidity and mortality.[22]

We present the derivation and validation of a new risk prediction algorithm for assessing the risk of

developing type 2 diabetes among a very large and unselected population derived from family practice, with appropriate weightings for ethnicity and social deprivation. We designed the algorithm (the QDScore) so that it would be based on variables that are readily available in patients' electronic health records or which patients themselves would be likely to know—that is, without needing laboratory tests or clinical measurements—thereby enabling it to be readily and cost effectively implemented in routine clinical practice and by national screening initiatives.

METHODS

Study design and data source

We did a prospective cohort study in a large population of primary care patients from version 19 of the QResearch database (www.qresearch.org). This is a large, validated primary care electronic database containing the health records of 11 million patients registered with 551 general practices using the Egton Medical Information System (EMIS) computer system. Practices and patients contained on the database are nationally representative for England and Wales and similar to those on other large national primary care databases using other clinical software systems.[23]

Practice selection

We included all QResearch practices in England and Wales once they had been using their current EMIS system for at least a year, so as to ensure completeness of recording of morbidity and prescribing data. We randomly allocated two thirds of practices to the derivation dataset and the remaining third to the validation dataset; we used the simple random sampling utility in Stata to assign practices to the derivation or validation cohort.

Cohort selection

We identified an open cohort of patients aged 25-79 years at the study entry date, drawn from patients registered with eligible practices during the 15 years between 1 January 1993 and 31 March 2008. We used an open cohort design, rather than a closed cohort design, as this allows patients to enter the population throughout the whole study period rather than requiring registration on a fixed date; our cohort should thus reflect the realities of routine clinical practice. We excluded patients with a prior recorded diagnosis of diabetes (type 1 or 2), temporary residents, patients with interrupted periods of registration with the practice, and those who did not have a valid postcode related Townsend deprivation score (about 4% of the population).

For each patient, we determined an entry date to the cohort, which was the latest of their 25th birthday, their date of registration with the practice, the date on which the practice computer system was installed plus one year, and the beginning of the study period (1 January 1993). We included patients in the analysis once they had a

Table 1 | Characteristics of patients aged 25-79 free of diabetes at baseline in derivation and validation cohorts between 1993 and 2008. Values are numbers (percentages) unless stated otherwise

	Derivation cohort		Validation cohort	
Characteristic	Women	Men	Women	Men
No of patients	1 283 135	1 257 618	622 488	610 344
Total person years' observation	8 373 101	8 063 034	3 898 407	3 744 630
No of incident cases of type 2 diabetes	34 916	43 165	16 912	20 623
Mean (SD) Townsend score	−0.19 (3.4)	−0.12 (3.4)	−0.15 (3.5)	−0.32 (3.6)
Median (interquartile range) age (years)	41 (31-56)	41 (32-54)	42 (32-56)	41 (32-54)
Ethnicity:				
White or not recorded	1 240 470 (96.67)	1 220 355 (97.04)	600 454 (96.46)	589 570 (96.60)
Indian	6 713 (0.52)	6 544 (0.52)	4 044 (0.65)	4 255 (0.70)
Pakistani	4 097 (0.32)	4 707 (0.37)	1 696 (0.27)	1 874 (0.31)
Bangladeshi	1 557 (0.12)	1 876 (0.15)	2 078 (0.33)	2 745 (0.45)
Other Asian	4 075 (0.32)	3 322 (0.26)	1 908 (0.31)	1 477 (0.24)
Black Caribbean	6 014 (0.47)	4 416 (0.35)	2 632 (0.42)	2 020 (0.33)
Black African	9 362 (0.73)	7 695 (0.61)	3 762 (0.60)	3 336 (0.55)
Chinese	2 619 (0.20)	1 709 (0.14)	1 435 (0.23)	948 (0.16)
Other, including mixed	8 228 (0.64)	6 994 (0.56)	4 479 (0.72)	4 119(0.67)
Risk factors:				
Ethnicity recorded	339 209 (26.44)	278 920 (22.18)	153 634 (24.68)	126 698 (20.76)
Body mass index recorded	1 013 326 (78.97)	895 308 (71.19)	498 397 (80.07)	440 159 (72.12)
Smoking recorded	1 154 858 (90.00)	1 046 823 (83.24)	566 602 (91.02)	514 693 (84.33)
Body mass index and smoking recorded	1 001 291 (78.03)	881 796 (70.12)	491 952 (79.03)	432 406 (70.85)
Family history of diabetes	148 466 (11.57)	102 583 (8.16)	68 500 (11.00)	47 569 (7.79)
Current smoker	298 455 (23.26)	349 294 (27.77)	149 492 (24.02)	173 076 (28.36)
Treated hypertension	74 436 (5.80)	60 232 (4.79)	39 174 (6.29)	32 131 (5.26)
Cardiovascular disease	32 447 (2.53)	51 601 (4.10)	16 975 (2.73)	27 222 (4.46)
Corticosteroids at baseline	22 424 (1.75)	14 738 (1.17)	12 721 (2.04)	8 190 (1.34)

Table 2 | Characteristics of men and women in derivation cohort with and without complete data for body mass index and smoking. Values are numbers (percentages) unless stated otherwise

	Body mass index				Smoking status			
	Women with missing data	Women with complete data	Men with missing data	Men with complete data	Women with missing data	Women with complete data	Men with missing data	Men with complete data
No of patients	269 809	1 013 326	362 310	895 308	128 277	1 154 858	210 795	1 046 823
Mean (SD) Townsend score	0.19 (3.5)	−0.3 (3.4)	0.21 (3.5)	−0.25 (3.4)	0.39 (3.5)	−0.26 (3.4)	0.4 (3.5)	−0.22 (3.4)
Mean (SD) age	45 (17)	45 (15)	42 (14)	44 (14)	46 (17)	45 (15)	42 (14)	44 (14)
Mean (SD) body mass index	NA	25 (4.8)	NA	26 (4)	26 (5)	25 (4.8)	26 (4.2)	26 (4)
Family history of diabetes	9 884 (3.66)	138 582 (13.68)	7497 (2.07)	95 086 (10.62)	2 189 (1.71)	146 277 (12.67)	2450 (1.16)	100 133 (9.57)
Current smoker	42 483 (15.75)	255 972 (25.26)	61 714 (17.03)	287 580 (32.12)	NA	298 455 (25.84)	NA	349 294 (33.37)
Treated hypertension	8 368 (3.10)	66 068 (6.52)	5 740 (1.58)	54 492 (6.09)	2 500 (1.95)	71 936 (6.23)	2 067 (0.98)	58 165 (5.56)
Cardiovascular disease	5 829 (2.16)	26 618 (2.63)	7 030 (1.94)	44 571 (4.98)	2 288 (1.78)	30 159 (2.61)	3 025 (1.44)	48 576 (4.64)
Corticosteroids at baseline	3 376 (1.25)	18 688 (1.84)	2 521 (0.70)	12 217 (1.36)	1 380 (1.08)	21 044 (1.82)	1 044 (0.50)	13 694 (1.31)
% (95% CI) observed risk of diabetes at 10 years	5.29 (5.15 to 5.44)	3.95 (3.89 to 4.00)	3.86 (3.76 to 3.96)	5.78 (5.71 to 5.85)	6.73 (6.46 to 7.01)	4.00 (3.95 to 4.06)	3.98 (3.84 to 4.13)	5.50 (5.44 to 5.56)

NA=not applicable.

minimum of one year's complete data in their medical record.[24] For each patient, we determined the right censor date, which was the earliest of the date of diagnosis of type 2 diabetes, date of death, date of deregistration with the practice, date of last upload of computerised data, or the study end date (31 March 2008).

Primary outcomes

Our primary outcome measure was the first (incident) diagnosis of type 2 diabetes mellitus as recorded on the general practice computer records. We identified patients with diabetes by searching the electronic health record for a diagnosis Read code for diabetes (C10%). As in other studies, we classified patients as having type 1 diabetes if they had a diagnosis of diabetes and had been prescribed insulin under the age of 35 and classified the remaining patients as having type 2 diabetes.[25]

Diabetes risk factors

We examined the following variables for inclusion in our analysis, all of which are known or thought to affect risk of developing diabetes,[16-19 26-29] and are also likely to be recorded in the patients' electronic records as part of routine clinical practice: self assigned ethnicity (nine categories); age at study entry (in single years); body mass index (continuous); smoking status (current smoker, not a current smoker); Townsend deprivation score (2001 census data evaluated at output areas as a continuous variable) ranging from −6 in the most affluent to 11 in the most deprived; recorded family history of diabetes in a first degree relative (binary variable yes/no); diagnosis of cardiovascular disease at baseline (binary variable yes/no); treated hypertension at baseline—that is, diagnosis of hypertension plus more than two prescriptions for antihypertensive drugs (binary variable yes/no); systemic corticosteroids at baseline—that is, at least two prescriptions within the preceding six months (binary variable yes/no).

We restricted all values of these variables to those that had been recorded in the person's electronic healthcare record before the diagnosis of type 2 diabetes (or before censoring for those who did not develop type 2 diabetes). We used Read codes for ethnicity to denote self assigned ethnicity. The Read classification is the coding system in use in general practice in England and Wales (ICD-10 is the equivalent coding system in use in hospitals). We grouped the codes into the English National Health Service standard 16+1 categories for the initial descriptive analysis. We then combined these 16+1 categories into the final nine reporting groups, thereby ensuring sufficient numbers of events in each group to enable a meaningful analysis. The "white or not recorded" category comprised British, Irish, and other white background, as well as those whose ethnicity was not recorded. We designated this as the reference category. We combined the group for whom ethnicity was not recorded with the white ethnic group; assuming the study population is comparable to the United Kingdom population, 93% or more of people without ethnicity recorded would be expected to be from a white ethnic group. The "other including mixed category" comprised "white and black Caribbean," "white and black African," "white and Asian," "other mixed," "other black, and other ethnic group." The "other Asian" category included Read codes for East African Asian, Indo-Caribbean, Punjabi, Kashmiri, Sri Lankan, Tamil, Sinhalese, Caribbean Asian, British Asian, mixed Asian, or Asian unspecified.

For body mass index and smoking status, we used the values recorded closest to the study entry date. We used body mass index rather than waist circumference, as the latter is not well recorded on clinical computer systems in the UK.

Model derivation and development

We calculated crude incidence rates of type 2 diabetes according to age, ethnic group, and deprivation in fifths. We then directly age standardised the incidence rates by ethnic group and deprivation by using the age distribution in five year bands of the entire derivation cohort as the standard population. We also used the same method to age standardise the means of

continuous variables and proportions with risk factors by ethnic group.

We used a Cox proportional hazards model in the derivation dataset to estimate the coefficients and hazard ratios associated with each potential risk factor for the first ever recorded diagnosis of diabetes for men and women separately. As in a previous study,[30] we used the Bayes information criterion to compare models.[31] This is a likelihood measure in which lower values indicate better fit and in which a penalty is paid for increasing the number of variables in the model. We used fractional polynomials to model non-linear risk relations with continuous variables where appropriate.[32 33] We tested for interactions between each variable and age and between smoking and deprivation and included significant interactions in the final model. Continuous variables were centred for analysis.

We used multiple imputation to replace missing values for smoking status and body mass index, and we used these values in our main analyses. We fitted our final model on the basis of multiply imputed datasets by using Rubin's rules to combine estimates of effects and standard errors of estimates to allow for the uncertainty caused by missing data.[34] Multiple imputation is a statistical technique designed to reduce the biases that can occur in "complete case" analysis along with a substantial loss of power and precision.[35-37] The imputation technique involves creating multiple copies of the data and replaces missing values with imputed values on the basis of a suitable random sample from their predicted distribution. Multiple imputation therefore allows patients with incomplete data to still be included in analyses, thereby making full use of all the available data, and thus increasing power and precision, but without compromising validity.[38] We used the ICE procedure in Stata to obtain five imputed datasets (further details are available from the corresponding author).[39]

We took the regression coefficient (that is, the log of the hazard ratio) for each variable from the final model and used these as weights for the new disease risk equations for type 2 diabetes. We combined these weights with the baseline survivor function for diagnosis of diabetes evaluated at 10 years and centred on the means of continuous risk factors to derive a risk equation for 10 years' follow-up. We have presented the Townsend coefficients in standard deviation units so that this can be applied in a non-UK setting where other indices of deprivation might apply.

We compared our final model (model A) with three other models in order to determine the additional contribution to the fit (using the Bayes information criterion in which lower values indicate better fit) and performance of the model of including both ethnicity and deprivation in the algorithm. Our first supplementary model (model B) included all the variables except for deprivation and ethnicity, the second model (model C) included deprivation but not ethnicity, and the third (model D) included ethnicity but not deprivation.

Validation of the QDScore

We tested the performance of the final algorithm (the QDScore) in the validation dataset. We calculated the 10 year estimated risk of acquiring type 2 diabetes for each patient in the validation dataset by using multiple imputations to replace missing values for smoking status and body mass index, as in the derivation dataset.

We calculated the mean predicted risk and the observed risk of diabetes at 10 years and compared these by 10th of predicted risk. The observed risk at 10 years was obtained by using the 10 year Kaplan-Meier estimate. We calculated the Brier score (a measure of goodness of fit where lower values indicate better accuracy[40]) by using the censoring adjusted

Table 3 | Crude and age standardised incidence of type 2 diabetes per 1000 person years by sex, deprivation fifth, and ethnicity in derivation dataset

	Women			Men		
	Person years	Crude rate	Age standardised rate (95% CI)	Person years	Crude rate	Age standardised rate (95% CI)
Townsend fifth						
1 (most affluent)	2 080 246	3.07	3.00 (2.93 to 3.08)	1 958 014	4.86	4.48 (4.39 to 4.57)
2	1 789 575	3.52	3.44 (3.35 to 3.52)	1 696 345	5.16	4.88 (4.78 to 4.98)
3	1 669 677	4.25	4.19 (4.09 to 4.29)	1 591 842	5.56	5.54 (5.43 to 5.66)
4	1 553 816	5.03	5.17 (5.05 to 5.28)	1 511 807	5.74	6.15 (6.02 to 6.28)
5 (most deprived)	1 259 406	5.81	6.39 (6.25 to 6.54)	1 286 781	5.73	6.56 (6.41 to 6.71)
Ethnicity						
White/not recorded	8 176 581	4.16	4.13 (4.08 to 4.17)	7 900 533	5.33	5.31 (5.26 to 5.36)
Indian	31 535	6.41	7.90 (6.73 to 9.08)	28 127	8.64	9.60 (8.35 to 10.85)
Pakistani	18 735	8.49	11.19 (9.16 to 13.21)	19 634	9.88	13.22 (11.24 to 15.21)
Bangladeshi	6 683	11.37	18.20 (12.93 to 23.47)	6 944	12.82	19.34 (14.28 to 24.4)
Other Asian	13 056	3.45	6.08 (2.73 to 9.44)	9 588	7.09	8.09 (6.03 to 10.15)
Caribbean	36 205	5.72	7.35 (6.28 to 8.43)	25 431	6.96	6.97 (5.89 to 8.05)
Black African	28 670	3.52	5.99 (4.54 to 7.44)	23 025	5.43	8.77 (6.84 to 10.7)
Chinese	9 547	3.35	5.40 (3.2 to 7.6)	6 603	3.33	3.32 (1.87 to 4.78)
Other	31 708	3.56	5.91 (4.51 to 7.3)	24 904	5.02	6.84 (5.51 to 8.18)

Table 4 | Distribution of risk factors for type 2 diabetes by ethnic group in men and women in derivation cohort. Values are age standardised means and proportions with 95% confidence intervals

	Mean Townsend score*	Mean body mass index	Percentage current smokers	Percentage family history of diabetes	Percentage treated hypertension	Percentage cardiovascular disease at baseline
Women						
White/not recorded	−0.28 (−0.29 to −0.27)	25.47 (25.46 to 25.48)	26.26 (26.18 to 26.34)	11.32 (11.27 to 11.38)	6.20 (6.16 to 6.24)	2.79 (2.77 to 2.82)
Indian	1.03 (0.95 to 1.12)	25.43 (25.3 to 25.55)	6.90 (6.23 to 7.57)	32.05 (30.87 to 33.22)	8.16 (7.3 to 9.02)	3.35 (2.72 to 3.98)
Pakistani	2.40 (2.3 to 2.5)	27.21 (27.02 to 27.41)	5.25 (4.55 to 5.95)	25.69 (24.25 to 27.12)	6.78 (5.62 to 7.93)	3.38 (2.52 to 4.23)
Bangladeshi	4.59 (4.43 to 4.76)	25.63 (25.34 to 25.92)	8.19 (6.61 to 9.78)	20.31 (18.36 to 22.26)	7.67 (5.74 to 9.6)	2.41 (1.14 to 3.69)
Other Asian	2.06 (1.91 to 2.2)	24.65 (24.44 to 24.87)	9.53 (8.46 to 10.59)	25.21 (23.63 to 26.79)	8.27 (6.69 to 9.85)	1.73 (0.89 to 2.58)
Black Caribbean	3.63 (3.55 to 3.71)	27.73 (27.59 to 27.87)	18.30 (17.34 to 19.27)	32.63 (31.41 to 33.85)	16.59 (15.55 to 17.62)	3.19 (2.6 to 3.79)
Black African	4.00 (3.92 to 4.09)	28.44 (28.29 to 28.58)	4.61 (4.05 to 5.18)	18.51 (17.44 to 19.58)	13.43 (12.22 to 14.65)	2.24 (1.54 to 2.94)
Chinese	2.24 (2.08 to 2.41)	22.87 (22.68 to 23.06)	7.08 (5.86 to 8.3)	15.07 (13.53 to 16.61)	7.30 (5.62 to 8.99)	2.06 (1.06 to 3.05)
Other	2.95 (2.85 to 3.05)	26.27 (26.12 to 26.42)	19.06 (18.1 to 20.02)	23.46 (22.34 to 24.59)	10.49 (9.41 to 11.58)	3.47 (2.69 to 4.26)
Men						
White/not recorded	−0.20 (−0.2 to −0.19)	26.15 (26.15 to 26.16)	33.49 (33.4 to 33.58)	8.07 (8.02 to 8.12)	5.28 (5.25 to 5.32)	4.54 (4.5 to 4.57)
Indian	1.11 (1.03 to 1.19)	25.24 (25.14 to 25.34)	22.71 (21.6 to 23.81)	29.95 (28.78 to 31.11)	9.13 (8.28 to 9.98)	6.68 (5.91 to 7.44)
Pakistani	2.43 (2.34 to 2.52)	25.74 (25.6 to 25.87)	32.82 (31.29 to 34.35)	24.42 (23.12 to 25.72)	5.96 (5.05 to 6.87)	7.07 (6.09 to 8.06)
Bangladeshi	4.38 (4.21 to 4.54)	24.51 (24.31 to 24.7)	46.04 (43.16 to 48.92)	20.20 (18.28 to 22.12)	6.60 (4.94 to 8.26)	9.70 (7.76 to 11.65)
Other Asian	2.32 (2.17 to 2.47)	25.26 (25.1 to 25.43)	28.11 (26.16 to 30.07)	20.56 (19.05 to 22.07)	6.80 (5.39 to 8.21)	5.40 (4.04 to 6.76)
Black Caribbean	3.72 (3.63 to 3.82)	26.22 (26.09 to 26.35)	40.45 (38.99 to 41.91)	24.65 (23.38 to 25.92)	11.09 (10.17 to 12.01)	3.48 (2.9 to 4.06)
Black African	4.10 (4 to 4.2)	26.05 (25.92 to 26.18)	17.95 (16.76 to 19.14)	13.78 (12.73 to 14.83)	12.02 (10.7 to 13.33)	2.64 (1.85 to 3.43)
Chinese	2.40 (2.21 to 2.6)	23.80 (23.59 to 24.02)	26.63 (24.23 to 29.03)	13.16 (11.33 to 14.99)	4.12 (2.57 to 5.67)	2.26 (1.15 to 3.37)
Other	3.12 (3.02 to 3.22)	25.95 (25.82 to 26.08)	35.18 (33.82 to 36.54)	18.65 (17.57 to 19.73)	7.25 (6.29 to 8.21)	3.61 (2.88 to 4.35)

*Measure of material deprivation, ranging from −6 (most affluent) to 11 (most deprived).

version adapted for survival data,[41] D statistic (a measure of discrimination where higher values indicate better discrimination),[42] and an R^2 statistic (a measure of explained variation for survival data, where higher values indicate that more variation is explained).[43] We also calculated the area under the receiver operator curve, where higher values indicate better discrimination. We also compared the performance of the QDScore with the Cambridge risk score,[16] which includes age, sex, body mass index, smoking status, corticosteroids, antihypertensive treatment, and family history of diabetes.

We calculated the proportion of patients in the validation sample who had an estimated 10 year risk of diagnosed diabetes of ≥10%, ≥15%, ≥20%, ≥30%, ≥40%, and ≥50% by age, sex, ethnic group, and deprivation according to the QDScore.

We used all the available data on the QResearch database and therefore did not do a pre-study sample size calculation. We used Stata (version 10) for all analyses and chose a significance level of 0.01 (two tailed).

RESULTS

Description of the derivation and validation dataset

Overall, 531 UK practices met our inclusion criteria, of which 355 were randomly assigned to the derivation dataset and 176 to the validation dataset. We excluded 20 practices: four practices had not completely uploaded all their electronic data for the relevant study period, seven practices were from Scotland, and nine practices were from Northern Ireland.

The derivation cohort contained 2 594 578 patients, of whom 53 825 had type 1 or type 2 diabetes before the start of the study and were therefore excluded leaving 2 540 753 patients (1 283 135; 50.50% women) aged 25-79 years and free of diabetes at baseline for analysis. The validation cohort contained 1 261 419 patients aged 25-79, of whom 28 587 had a previous diagnosis of type 1 or type 2 diabetes leaving 1 232 832 patients for analysis (50.49% women).

Overall, we studied 3 773 585 patients contributing 24 079 172 person years, of whom 115 616 patients (78 081 in the derivation cohort and 37 535 in the validation cohort) had a new diagnosis of type 2 diabetes during follow-up. Table 1 compares the characteristics of eligible patients in the derivation and validation cohorts. Although this validation cohort was drawn from an independent group of practices, the baseline characteristics were very similar to those for the derivation cohort. Overall, 898 461 patients (23.81% of 3 773 585) had ethnicity recorded, and 122 736 (13.66%) of these were from a non-white ethnic group. Practices in areas where the proportion of patients from a non-white ethnic group is higher according to the 2001 census (such as London (28.9%), East Midlands (6.5%), and West Midlands (11.3%)) also have higher rates of completeness of recording of ethnicity on the QResearch database (40.1%, 21.4%, and 30.1% for the above areas).

Patterns of missing data

Table 1 shows that 78.97% of women in the derivation cohort had body mass index recorded and 90.00% had smoking status recorded; 78.03% had both body mass index and smoking status recorded. For men, the corresponding figures were 71.19%, 83.24%, and 70.12%. Overall, 22.97% of women and 29.88% of men had either smoking or body mass index imputed by multiple imputation (data were not imputed for ethnicity—all patients with missing ethnicity were treated as white/not recorded). Similar figures were observed for men and women in the validation cohort, where multiple imputation was also used.

Table 2 shows the characteristics of men and women with complete data for smoking and body mass index compared with those who had missing data. Women with missing data had different patterns of risk factors—for example, women with complete data for body mass index were more likely to have a family history of diabetes, to be recorded as current smokers, and to have treated hypertension. They also had a lower 10 year observed risk of diabetes compared with women with missing body mass index data. Women with complete data for smoking were more likely to have a diagnosis of cardiovascular disease, a diagnosis of treated hypertension, and a family history of diabetes. The 10 year observed risk of diabetes was lower than for women whose smoking status was missing. The pattern was similar for men for most risk factors, except that the observed risks of diabetes were lower among men with missing data.

Table 5 | Adjusted hazard ratios (95% confidence interval) for QDScore in derivation cohort (see fig 1 for graphical representation of interaction terms)

	Women	Men
White/not recorded	1	1
Indian	1.710 (1.488 to 1.965)	1.929 (1.700 to 2.189)
Pakistani	2.152 (1.839 to 2.517)	2.538 (2.202 to 2.925)
Bangladeshi	4.071(3.242 to 5.112)	4.532 (3.673 to 5.591)
Other Asian	1.264 (0.943 to 1.695)	1.894 (1.492 to 2.404)
Black Caribbean	0.798 (0.695 to 0.915)	0.955 (0.824 to 1.108)
Black African	0.805 (0.661 to 0.979)	1.695 (1.421 to 2.023)
Chinese	1.961 (1.385 to 2.777)	1.414 (0.928 to 2.154)
Other	0.889 (0.738 to 1.07)	1.199 (1.005 to 1.431)
Age 1†	84.059 (68.345 to 103.384)	105.666 (89.11 to 125.3)
Age 2‡	0.995 (0.9946 to 0.9954)	0.996 (0.9955 to 0.9962)
BMI 1§	37.293 (31.118 to 44.694)	3.168 (3.000 to 3.345)
BMI 2¶	0.934 (0.928 to 0.939)	0.832 (0.822 to 0.841)
Townsend score (per increase of 1 SD)	1.201 (1.188 to 1.214)	1.140 (1.129 to 1.152)
Family history of diabetes in a first degree relative	2.358 (2.278 to 2.441)	2.725 (2.638 to 2.815)
Current smoker	1.268 (1.225 to 1.312)	1.249 (1.214 to 1.285)
Treated hypertension	1.787 (1.738 to 1.837)	1.711 (1.665 to 1.759)
Diagnosis of cardiovascular disease	1.458 (1.402 to 1.517)	1.500 (1.455 to 1.546)
Current treatment with corticosteroids	1.412 (1.339 to 1.489)	1.259 (1.181 to 1.342)

Model also included fractional polynomial terms for age and body mass index and interactions between age terms and body mass index terms, age terms and family history of diabetes, and age terms and smoking status (see fig 1).
BMI=body mass index.
†Women: age 1=$(age/10)^{½}$; men: age 1=log(age/10).
‡Women: age 2=$(age/10)^3$; men: age 2=$(age/10)^3$.
§Women: BMI 1=(BMI/10); men: BMI 1=$(BMI/10)^2$.
¶Women: BMI 2=$(BMI/10)^3$; men: BMI 2=$(BMI/10)^3$.

Incidence of diabetes

Table 3 shows the crude and age standardised rates of type 2 diabetes by sex, deprivation, and ethnicity in the derivation cohort. The age standardised rates for the white reference group were 4.13 (95% confidence interval 4.08 to 4.17) per 1000 person years for women and 5.31 (5.26 to 5.36) per 1000 person years for men. The crude and age standardised incidence rates of type 2 diabetes in the derivation cohort varied widely between ethnic groups, as shown in table 3. Age standardised rates were significantly higher for men in every ethnic group compared with the white reference group, except for Chinese men. In women, age standardised incidence rates were higher for every group compared with the white reference group. The highest age standardised rates were in South Asians, and significant differences existed between the South Asian groups. For example, the rate for Bangladeshi women was 18.20 (12.93 to 23.47) per 1000 person years and that for Bangladeshi men was 19.34 (14.28 to 24.4) per 1000 person years. For Pakistanis, the corresponding rates per 1000 person years were 11.19 (9.16 to 13.21) for women and 13.22 (11.24 to 15.21) for men.

We also found a marked difference in the age standardised incidence rates of type 2 diabetes by deprivation, with a more than twofold difference for women when comparing the most deprived fifth (6.39 (6.25 to 6.54) per 1000 person years) with the most affluent fifth (3.00 (2.93 to 3.08) per 1000 person years). A similar, but less steep gradient was seen for men. The rates seen in the validation cohort were similar to those for the derivation cohort (data not shown).

Prevalence of risk factors by ethnicity

Table 4 shows the age standardised distribution of risk factors across each of the main ethnic groups. Substantial heterogeneity exists across the ethnic groups for risk factors, and the distribution also differs between men and women within ethnic groups. The notable results include substantial differences in the age standardised prevalence of smoking among men of Bangladeshi (46.04%, 95% confidence interval 43.16% to 48.92%), Caribbean (40.45%, 38.99% to 41.91%), Pakistani (32.82%, 31.29% to 34.35%), white/not recorded (33.49%, 33.40% to 33.58%), Chinese (26.63%, 24.23% to 29.03%), Indian (22.71%, 21.60% to 23.81%), and black African (17.95%, 16.76% to 19.14%) origin. Smoking rates were lower for women in each ethnic group compared with men but varied widely between women from different groups.

Treated hypertension was highest among black Caribbean and black African men and women and more than twice as high as that for the white reference group. Recorded family history of diabetes was highest among black Caribbean women (32.63%, 31.41% to

33.85%) and Indian men (29.95%, 28.78% to 31.11%), which was more than three times that for the white reference group who had the lowest rates (11.32%, 11.27% to 11.38% for women and 8.07%, 8.02% to 8.12% for men).

Bangladeshi men and women had the highest age standardised mean deprivation scores, followed by those of black African and black Caribbean origin. Indians and the white reference group had the lowest mean deprivation scores, as shown in table 4.

The highest mean body mass index was seen among black African women (age standardised mean 28.44, 28.29 to 28.58) compared with 25.47 (25.46 to 25.48) for women in the white reference group. The lowest value was in Chinese women (age standardised mean 22.87, 22.68 to 23.06). Similar patterns, although slightly less marked, were seen for men across the ethnic groups. Finally, 9.70% (7.76% to 11.65%) of Bangladeshi men had a recorded diagnosis of cardiovascular disease at baseline, which was more than twice that for men in the white reference group (4.54%, 4.50% to 4.57%) and more than four times that found in Chinese men (2.26%, 1.15% to 3.37%).

Model development

Table 5 shows the results of the Cox regression analysis for the QDScore. After adjustment for all other variables in the model, we found significant associations with risk of type 2 diabetes in both men and women for age, body mass index, family history of diabetes, smoking status, treated hypertension, use of corticosteroids, diagnosed cardiovascular disease, social deprivation, and ethnicity. We therefore included these variables in the final model and risk prediction algorithm.

We found significant heterogeneity of risk of type 2 diabetes by ethnic group compared with the white reference population, having adjusted for age, body mass index, deprivation, family history of diabetes, smoking status, treated hypertension, diagnosed

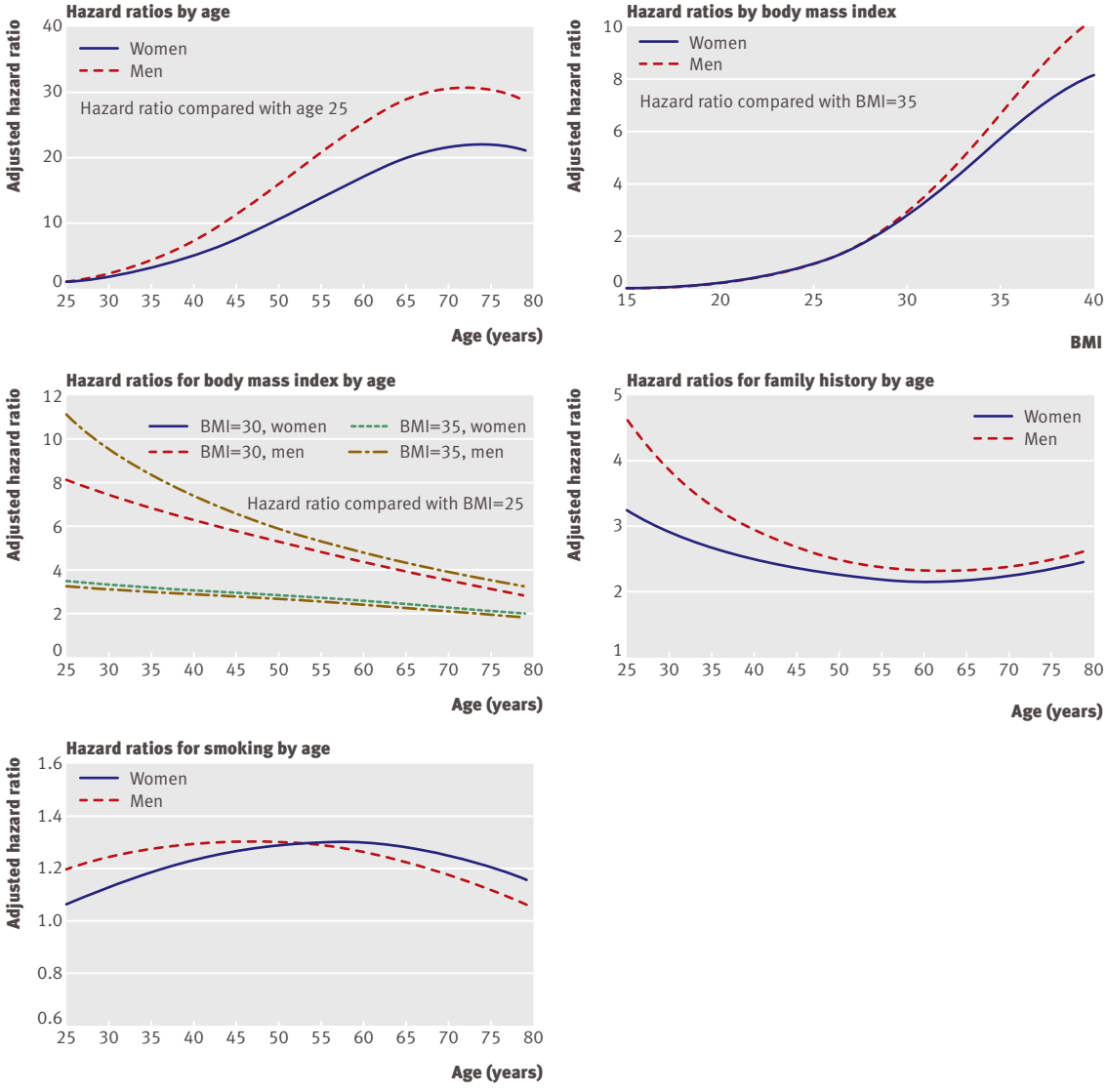

Fig 1 | Graphical representation of age interactions for men and women for risk of type 2 diabetes

Table 6 | Validation statistics for QDScore and Cambridge risk score in validation cohort. Values are mean (95% confidence interval)

	QDScore	Cambridge risk score
Women		
R squared*	51.53 (50.90 to 52.16)	45.77 (45.08 to 46.46)
D statistic*	2.110 (2.084 to 2.137)	1.880 (1.854 to 1.906)
ROC statistic*	0.853 (0.850 to 0.856)	0.813 (0.810 to 0.817)
Brier score†	0.058 (0.055 to 0.060)	0.044 (0.041 to 0.046)
Men		
R squared*	48.16 (47.52 to 48.80)	41.82 (41.19 to 42.51)
D statistic*	1.973 (1.947 to 1.998)	1.735 (1.710 to 1.760)
ROC statistic*	0.834 (0.831 to 0.836)	0.801 (0.798 to 0.804)
Brier score†	0.078 (0.075 to 0.080)	0.055 (0.052 to 0.057)

ROC=receiver operator curve.
*Higher values indicate better discrimination.
†Lower values indicate better performance.

cardiovascular disease, use of corticosteroids, and diagnosed cardiovascular disease, as shown in table 4. For example, among Bangladeshis, the adjusted hazard ratio for women was 4.07 (95% confidence interval 3.24 to 5.11) and that for men was 4.53 (3.67 to 5.59). These were significantly higher than the increased hazard ratios in Pakistani women and men (2.15, 1.84 to 2.52; and 2.54, 2.20 to 2.93). Both Pakistani and Bangladeshi men had significantly higher hazard ratios than Indian men. Black African men and Chinese women had increased risks compared with the corresponding white reference group. The only groups to have significantly lower risks than the white reference group were black African women (0.81, 0.66 to 0.98) and black Caribbean women (0.80, 0.70 to 0.92).

The fractional polynomial terms selected for inclusion in the model were as follows. For age in women the two terms were $(age/10)^{½}$ and $(age/10)^3$. For body mass index in women, the two terms were (bmi/10) and $(bmi/10)^3$. For men, the two age terms were log(age/10) and $(age/10)^3$ and the two terms for body mass index were $(bmi/10)^2$ and $(bmi/10)^3$. Figure 1 shows the estimated adjusted hazard ratios by age and body mass index for these fractional polynomial terms in men and women.

We identified significant interactions between age and body mass index, age and family history of diabetes, and age and smoking status. We therefore included these interactions in the final model, and the general direction of the effects was that body mass index and family history of diabetes tended to have a greater impact on risk of diabetes at younger ages, as shown in fig 1. Smoking had a more complex relation with age; the risk peaked in middle age for both men and women.

In a comparison of models, the median Bayes information criterion for women for our final model (model A) was 875 203, for the model without deprivation and ethnicity (model B) it was 876 400, for the model without ethnicity (model C) it was 875 270, and for the model without deprivation (model D) it was 876 198, indicating that the model that incorporated both ethnicity and deprivation was superior to the other three. For men, the corresponding figures were 1 086 755, 1 087 745, 1 087 034, and 1 087 369, similarly supporting the inclusion of both ethnicity and deprivation into the final model.

Calibration and discrimination of QDScore

Table 6 shows the results for the validation statistics for men and women after application of the QDScore and the Cambridge risk score in the validation dataset. The QDScore shows higher levels of discrimination than the Cambridge risk score. For example, in women the D statistic for the QDScore was 2.11 (95% confidence interval 2.08 to 2.14) compared with 1.88 (1.85 to 1.91) with the Cambridge risk score; a 0.1 difference in the D statistic indicates an important difference in prognostic separation between two risk algorithms.[41] The QDScore explained a higher proportion of the variation—it explained 51.53% of the variation in women and 48.16% of that in men. The corresponding values for the Cambridge risk score were 45.77% and 41.82%. The Brier score, however, was slightly lower for the Cambridge risk score in both men and women.

Figure 2 compares the mean predicted scores from the QDScore with the observed risks at 10 years within each 10th of predicted risk in order to assess the calibration of the model in the validation sample. The close correspondence between predicted and observed 10 year risks within each model 10th suggests that the model was well calibrated. For example, in the top 10th of risk, the mean predicted risk was 18.31% (95% confidence interval 18.24% to 18.38%) in women and the observed risk was 18.82% (18.39% to 19.26%). The ratio of predicted to observed risk in this tenth was 0.97, indicating almost perfect calibration (a ratio of 1 indicates perfect calibration—that is, no under-prediction or over-prediction). We found similar results for men, with a ratio of 0.99 in the top 10th of predicted risk.

Predictions with age, sex, deprivation, and ethnicity

Table 7 shows the percentages of men and women in the validation dataset with a 10 year predicted risk of being diagnosed as having type 2 diabetes according to a range of thresholds and by age band. For example, at the 10% threshold, 10.60% of women and 15.06% of men had a 10% or higher predicted risk of being diagnosed as having type 2 diabetes over 10 years. This varied markedly by age such that 21.43% of women aged 55-59 and 30.99% of women aged 65-69 had a 10% or greater risk of being diagnosed as having type 2 diabetes over 10 years. The corresponding figures for men were 33.28% and 44.08%.

Tables 8 and 9 show the 10 year risk of type 2 diabetes among men and women of different ethnic groups and for those living in the most deprived and affluent areas. For example, 33.83% of Bangladeshi women had a 10 year risk of being diagnosed as having diabetes of 10% or more compared with 10.48% of women in the white reference group, and 15.03% of women in the most deprived fifth had a 10% or higher

risk of developing diabetes over the next 10 years compared with 6.52% of women in the most affluent fifth. The difference between affluent and deprived fifths is more marked for women than for men; the corresponding figures are 15.65% for men in the most deprived fifth and 13.21% for men in the most affluent fifth.

Overall, almost half (15 545/32 450; 47.9%) of cases of diabetes occurred in the top 10th of the distribution (risk of ≥10.38%) and almost 70% (22 476/32 450) occurred in the top fifth (risk of ≥5.98%).

DISCUSSION

The QDScore is the first diabetes prediction algorithm developed and validated by using routinely collected data to predict the 10 year risk of developing type 2 diabetes. Our final model includes both deprivation and ethnicity as well as age, sex, smoking, treated hypertension, body mass index, family history of diabetes, current treatment with corticosteroids, and previous diagnosis of cardiovascular disease. The QDScore does not require any laboratory testing or clinical measurements and so can be used in many settings, including by individual members of the public who have access to a computer. This risk prediction tool might be used to identify and proactively intervene in people identified as having an increased risk. This algorithm, like other algorithms that predict cardiovascular disease,[30 44] relies on routinely collected data and has the advantage that it is readily implementable. Furthermore, it is likely to reduce, rather than exacerbate, widespread and persistent health inequalities. The QDScore performed well compared with the Cambridge risk score. Assuming that the effectiveness and cost effectiveness of suitable interventions shown in randomised controlled trials extend to unselected patients from primary care,[7-10] the QDScore could be used to identify patients at increased risk of diabetes who might benefit from interventions to reduce their risk.

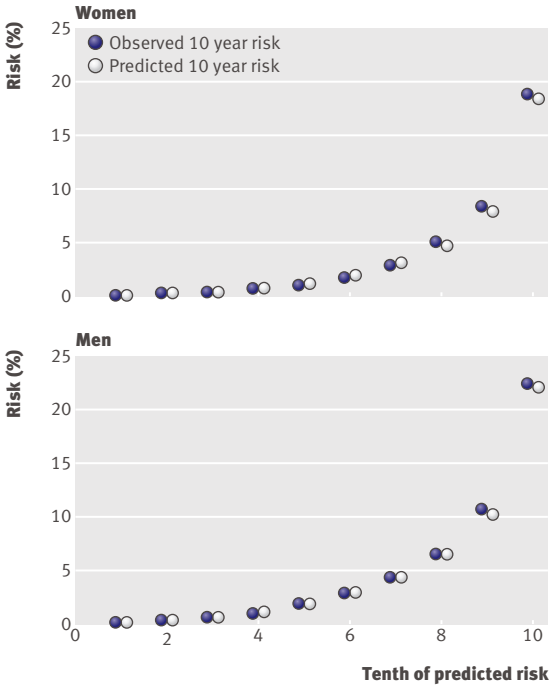

Fig 2 | QDScore predicted and observed risk of diabetes by 10th of predicted risk

The traditional method for identifying patients at increased risk of type 2 diabetes has involved the detection of impaired glucose tolerance requiring an inconvenient and expensive oral glucose tolerance test. Targeted screening of higher risk groups has been proposed as a more cost effective solution,[45] as the risk factors for diabetes and cardiovascular outcomes overlap considerably.[46] Less expensive and more practical methods of identifying patients at increased risk are needed; these should ideally be based on models developed from contemporaneous data in ethnically and socioeconomically diverse populations obtained from the clinical setting in which these models will subsequently be applied. Simple clinical models using readily available data can offer similar discrimination to more complex models using laboratory data or biomarkers,[17] and clinical models that do not need clinical measurements may have a further utility in settings where clinical measurements are not available or are too costly to collect.[47] UK datasets derived from family practices have the advantage of having large and broadly representative populations with historical data tracking back well over a decade in most practices. These databases also contain data on many of the key variables known to be associated with risk of type 2 diabetes, such as age, sex, ethnicity,[28 48 49] smoking,[16 24 50] body mass index,[16 17 28 48] family history of diabetes,[16 17 28 48 49] treated hypertension,[16 17] current use of corticosteroids,[16] and social deprivation.[51] Deprivation is not only strongly associated with increased prevalence of diabetes and diabetes related risk factors such as diet, obesity, and smoking but is also associated with poorer outcomes and intermediate measures such as achievement of lipid targets.[51]

Strengths and weakness

Sampling and generalisability

Particular strengths of our study are the use of a large representative population from a validated database, our prospective cohort design, and the substantial numbers of patients with self assigned ethnicity for use in the analysis. We have modelled interactions with age and included these in the final model, so our algorithm takes account of the differential effect of three key variables (family history of diabetes, body mass index, and smoking) at different ages.

Another important strength of the QDScore is that all the variables used in the algorithm will either be known to an individual patient or are collected as part of routine clinical practice and recorded within an individual patient's primary healthcare record in most economically developed countries. This means that the algorithm can be used by patients for self assessment in a web based calculator (www.qdscore.org) similar to

the one available for self assessment of cardiovascular disease (www.qrisk.org). Alternatively, it can be implemented within clinical computer systems used in primary care and be used to stratify the practice population (aged 25-79) for risk on a continuing basis without the need for manual entry of data. Although no widely agreed thresholds for classification of patients at "high risk" exist, the QDScore could act as a basis for a systematic programme to identify patients at increased risk for intervention or to aid earlier diagnosis. Importantly, appropriate weighting for ethnicity and social deprivation should furthermore help to avoid widening health inequalities associated with introduction of systematic programmes of disease prevention activities.

Limitations compared with an ideal study

Despite its strengths, our study has limitations compared with an ideal study. In the ideal study, a large representative and tightly phenotyped primary care cohort would be assembled and followed longitudinally over the course of a decade. No patients would be lost to follow-up, and all patients would be subjected to repeated oral glucose tolerance tests throughout follow-up to confirm or refute the diagnosis of type 2 diabetes. Although such a study would be very welcome, it would take at least 15 years to carry out and report, it would be unlikely to be feasible in routine primary care, and calibration would be inaccurate in more socially and ethnically diverse populations with different baseline risks. Our study offers a practical alternative approach, which can be implemented into primary care in a cost effective manner, while acknowledging the potential biases and their likely impact.

Potential sources of misclassification, bias, and confounding

One limitation of our study is that the main outcome was type 2 diabetes diagnosed by a clinician and recorded on the clinical computer system. The outcome was not formally validated, and we have not used the results of laboratory tests to confirm the diagnosis. However, this diagnosis would be unlikely to be recorded if the patient did not have diabetes—other studies of similar databases have shown good levels of accuracy for common chronic conditions, especially those that are now included in the UK quality and outcomes framework.[52]

Undiagnosed diabetes is a well recognised problem and is not specifically considered by our study. It is estimated to affect approximately 3% of the population according to the health survey for England.[53] Some evidence suggests that South Asian women are more likely to have undiagnosed diabetes than are the

Table 7 | Percentage of patients in validation dataset with 10 year predicted risk of type 2 diabetes from QDScore of ≥10%, ≥15%, ≥20%, ≥30%, ≥40%, and ≥50% by age and sex

	Predicted 10 year risk score for type 2 diabetes					
Age band (years)	≥10%	≥15%	≥20%	≥30%	≥40%	≥50%
Women						
25-29	0.34	0.10	0.02	0.00	0.00	0.00
30-34	0.93	0.36	0.12	0.02	0.01	0.00
35-39	2.40	1.03	0.47	0.07	0.02	0.01
40-44	5.15	2.19	1.09	0.29	0.09	0.03
45-49	9.52	4.29	2.32	0.70	0.21	0.08
50-54	15.24	7.42	4.00	1.31	0.47	0.17
55-59	21.43	11.39	6.19	2.14	0.81	0.28
60-64	27.79	15.20	8.63	3.22	1.31	0.52
65-69	30.99	16.49	9.20	3.03	1.17	0.45
70-74	30.70	15.68	8.25	2.62	0.89	0.29
75-79	27.75	12.38	5.94	1.71	0.50	0.17
Total	10.60	5.28	2.82	0.92	0.33	0.12
Men						
25-29	0.54	0.26	0.15	0.05	0.02	0.01
30-34	1.62	0.67	0.36	0.14	0.05	0.02
35-39	4.17	1.77	0.94	0.39	0.13	0.04
40-44	8.67	4.11	2.08	0.76	0.31	0.10
45-49	15.29	7.46	3.87	1.39	0.58	0.24
50-54	23.84	12.46	6.77	2.39	1.01	0.43
55-59	33.28	17.82	10.21	3.83	1.60	0.63
60-64	40.41	22.83	13.32	4.74	1.89	0.79
65-69	44.08	24.99	14.40	5.03	1.98	0.79
70-74	44.00	23.58	12.94	4.14	1.43	0.50
75-79	39.54	18.86	9.23	2.57	0.73	0.22
Total	15.06	7.89	4.36	1.53	0.60	0.23

Table 8 | Percentage of patients in validation dataset with 10 year predicted risk of type 2 diabetes from QDScore of ≥10%, ≥15%, ≥20%, ≥30%, ≥40%, and ≥50% by ethnicity and sex

	Predicted 10 year risk score for type 2 diabetes					
Ethnic group	≥10%	≥15%	≥20%	≥30%	≥40%	≥50%
Women						
White/not recorded	10.48	5.15	2.70	0.85	0.29	0.10
Indian	14.56	9.35	5.86	2.72	1.29	0.72
Pakistani	26.36	18.16	12.91	6.43	3.36	1.77
Bangladeshi	33.83	25.07	18.72	10.92	6.35	3.71
Other Asian	6.45	3.77	2.20	1.05	0.47	0.16
Black Caribbean	17.10	11.09	7.45	2.93	1.41	0.49
Black African	8.96	4.81	2.68	0.74	0.27	0.08
Chinese	4.32	2.58	1.46	0.28	0.00	0.00
Other	7.90	4.26	2.39	0.74	0.29	0.13
Men						
White/not recorded	14.93	7.72	4.21	1.43	0.54	0.20
Indian	20.52	13.07	8.55	4.21	2.09	1.20
Pakistani	27.21	19.32	13.93	7.42	3.68	2.19
Bangladeshi	33.26	24.52	19.23	11.58	7.10	4.26
Other Asian	14.96	9.75	6.30	2.71	1.35	0.74
Black Caribbean	19.31	11.98	7.72	3.47	1.63	0.79
Black African	16.10	9.83	6.15	3.24	1.83	1.02
Chinese	7.59	3.69	2.11	0.84	0.42	0.11
Other	9.91	6.00	3.69	1.72	0.80	0.27

general population, so our hazard ratios might underestimate the association for these patients.[54] The risk factors for diagnosed diabetes are very similar to those for undiagnosed diabetes.[55] Nevertheless, most previously undiagnosed cases are likely to have been included in the identified high risk groups and to have been picked up by systematic further evaluation, because risk stratification improves yield.[55]

Our study might have been affected by recording bias if a patient diagnosed as having diabetes was not recorded as having diabetes on the practice computer system. The recording bias could lead to misclassification of patients either at baseline or at follow-up and is part of the justification for having a targeted approach. Any misclassification bias of the outcome, if non-differential, would tend to bias the hazard ratio towards one and reduce discrimination.

Recording of a positive family history of diabetes was higher among women than among men. This could reflect recording bias or information biases resulting from differences in family history among women or greater opportunity for the information to be recorded as women tend to have higher consultation rates than men. Our study might have been affected by an ascertainment bias caused by differential testing of patients for diabetes by ethnic group or in those with specific risk factors. This could lead to increased rates of detection among patients with specific risk factors, including South Asian ethnicity, a family history of diabetes, or obesity—increased awareness among patients and clinicians might increase the likelihood of testing and therefore of clinical diagnosis. The effect of this would be to increase the apparent strength of the association between the risk factors and incident diabetes. Nonetheless, our hazard ratios for the risk factors in the model are generally of a similar magnitude to those found in other studies which tested for diabetes in the entire study cohort.[56] In addition, the assessment and recording of these factors in clinical practice is becoming increasingly routine and complete, so limiting the effect of this potential bias.

Another potential limitation of our study is that 25% of patients had missing values for either body mass index or smoking status. Patients with complete data tended to have different risks than those with missing data. We therefore used the technique of multiple imputation to substitute missing values for smoking and body mass index, rather than excluding these patients, as this is a less biased approach that makes the most efficient use of available data. The differences in risk factors and in the observed risks of diabetes between patients with and without missing data support the use of multiple imputation rather than a complete case analysis.

Variables included in final model

Clinicians had recorded our predictor variables on the clinical computer system before the diagnosis of type 2 diabetes, so these will not have been subject to recall bias. We have used the entire population registered with the QResearch practices contributing to the database from England and Wales. Consequently, the population is unlikely to be affected by selection bias, in contrast to the selection bias that inevitably occurs when patients are individually recruited to clinical cohorts or clinical trials.[57] We have included a proxy measure of material deprivation, the Townsend score,[58] which is based on the patient's postcode at

the level of the output area (corresponding to around 150 households) and is a composite score comprising lack of a car, unemployment, over-crowding, and non-home ownership. Some people living within an output area will not be typical of the other residents, resulting in some misclassification. Deprivation is likely to be associated with other factors known to increase risk of diabetes, such as poor diet, lack of exercise, and increased alcohol intake, and so will account at least in part for some of the effect of these factors.[51] Lastly, we included treated hypertension as a predictor as both blood pressure and some antihypertensive drugs (such as thiazides) may have contributed to the increased hazard ratios associated with this variable.

Self assigned ethnicity

We used self assigned ethnicity in our analyses, as reported by patients to their general practices, which has advantages over analyses in which ethnicity is assigned by an informant rather than the patient, is imputed geographically, or is related to country of birth. The last of these is particularly problematic as increasing numbers of people from minority ethnic groups are now being born in the UK. We have also been able to disaggregate the South Asian groups and report on them separately, which answers concerns with studies that tend to combine them into one group when they differ in risk factor exposure, disease rates, and outcomes. One important limitation is that only one quarter of patients overall had self assigned ethnicity recorded. Among those with a recorded value, 13.66% were recorded as from a minority ethnic group, which is higher than the estimated figure for 2006 based on the 2001 census, indicating over-representation of practices from ethnically diverse areas, that practices in ethnically diverse areas are more likely to record ethnicity, or most likely a combination of both. We have assumed that where patients have self assigned ethnicity recorded (as Bangladeshi, for example) this is accurate and the patient was indeed Bangladeshi. Where patients did not have ethnicity recorded, we have assumed they were white. Any misclassification arising from these assumptions is most likely to affect the reference category of "white or not recorded," but because of the mix of the populations of England and Wales, less than 7% of such patients are likely to be from a non-white ethnic group. This misclassification error is likely to be non-differential and if so will tend to underestimate the relative effect of ethnicity on risk of type 2 diabetes rather than generating spurious associations. Misclassification would also tend to reduce levels of discrimination and underestimate risk in some misclassified patients. We restricted all values of variables in the model to those that had been recorded in the person's electronic healthcare record before the diagnosis of type 2 diabetes (or before censoring for those who did not develop type 2 diabetes) in order to avoid recording bias.

Validation of risk prediction algorithm

We validated the QDScore in a separate sample of general practices from those used to develop the score. The QDScore has good discrimination (that is, ability to separate out people who did and those who did not subsequently develop type 2 diabetes) and explains approximately 50% of the total variation in times to diagnosis of diabetes. The D statistic, which is a measure of discrimination appropriate for survival type data, was higher than in our cardiovascular disease algorithm and that reported in some other studies.[30 42] This increases the likelihood that the algorithm will more accurately predict risk for an individual patient. An important limitation of our validation is that a degree of over-optimism could exist as, although we have used a completely physically discrete set of general practices for the validation, these practices use the same clinical computer system (EMIS) as those used to derive the algorithm. This system is, however, currently in use in 60% of UK general practices, so the diabetes clinical risk algorithm is at least likely to perform well for well over half of the UK's population. A more stringent test of performance would involve practices using a different clinical computer system;

Table 9 | Percentage of patients in validation dataset with 10 year predicted risk of diabetes from QDScore of ≥10%, ≥15%, ≥20%, ≥30%, ≥40%, and ≥50% by deprivation fifth and sex

	Predicted 10 year risk score for type 2 diabetes					
Townsend fifth	≥10%	≥15%	≥20%	≥30%	≥40%	≥50%
Women						
1 (most affluent)	6.52	2.64	1.24	0.28	0.09	0.02
2	8.29	3.61	1.71	0.43	0.13	0.03
3	10.83	5.05	2.49	0.72	0.22	0.08
4	12.94	6.76	3.61	1.21	0.40	0.14
5 (most deprived)	15.03	8.74	5.26	2.03	0.85	0.34
Men						
1 (most affluent)	13.21	6.17	3.07	0.93	0.32	0.10
2	14.80	7.30	3.71	1.21	0.40	0.13
3	15.99	8.31	4.45	1.48	0.53	0.19
4	15.91	8.73	4.95	1.78	0.74	0.29
5 (most deprived)	15.65	9.08	5.65	2.27	1.01	0.46

WHAT IS ALREADY KNOWN ON THIS TOPIC

Good evidence shows that behavioural or pharmacological interventions can prevent type 2 diabetes in up to two thirds of patients at high risk and that early diagnosis is likely to improve outcomes

In 2009 the Department of Health will start a major vascular screening programme, which includes identification and management of patients at high risk of diabetes for preventive care

No widely accepted and validated risk prediction score takes account of both social deprivation and ethnicity and can be applied in primary care in the UK

WHAT THIS STUDY ADDS

The QDScore is a new risk prediction algorithm for type 2 diabetes developed in a very large and unselected family practice derived population, with appropriate weightings for ethnicity and social deprivation

The final algorithm includes self assigned ethnicity, age, sex, body mass index, smoking status, family history of diabetes, Townsend deprivation score, treated hypertension, cardiovascular disease, and current use of corticosteroids

The performance of the QDScore in an independent sample of practices showed good discrimination and calibration

however, recording of ethnicity in other general practice databases is at present likely to be too low for a meaningful comparison, as EMIS has more practices in ethnically diverse areas. Nonetheless, our previous algorithm for cardiovascular disease, developed with similar methods and the same database,[44] has subsequently performed well on another database containing primary care data from practices using a different clinical computer system.[23]

Heterogeneity of risk factors and risk of type 2 diabetes

Our study has good face validity, as the prevalence of established risk factors reported here corresponds to that reported elsewhere.[59] We found a significant heterogeneity of risk factors, incidence rates, and hazard ratios for type 2 diabetes across the ethnic groups. The high prevalence of a recorded family history among South Asians may reflect a true increased rate or could be due to differences in what constitutes a first degree relative (for example, cousins may be regarded as siblings). Of particular interest are the significant differences in hazard ratios between the South Asian groups; Bangladeshi men and women had higher risks than Pakistanis, who in turn had higher risks than Indians.

Comparison with other diabetes risk scores

Routinely collected data from electronic primary healthcare records have been used to develop other risk prediction algorithms. For example, data from 531 general practices was used to develop and validate the QRISK2 cardiovascular disease risk tool, which is being implemented in clinical settings in the UK.[30 44] The Cambridge diabetes risk score was developed by combining data from two different general practice samples. The first sample consisted of half of the participants recruited for the study, in which patients were tested for diabetes by using an oral glucose tolerance test in one general practice in Cambridgeshire. The second sample consisted of half of the incident cases of diabetes identified over a 12 month period from 41 practices in the south of England.[16 20] The combined data from a total of 650 patients, including 126 cases of diabetes, were then used to derive a risk score designed to identify patients with undiagnosed diabetes at a point in time.[16 20] The score was then validated in the remaining half of the recruited patients from the practice in Cambridgeshire. The Cambridge risk score has since been applied to a prospective cohort to estimate the risk of incident diabetes in 25 000 people from Norfolk.[60]

One advantage of the QDScore is the use of a larger and more representative cohort, which is more likely to generalise to the UK. Another advantage is the inclusion of both deprivation and self assigned ethnicity, which are independently associated with risk of incident diabetes; this is likely to help with the problems identified with the Cambridge risk score in its performance in ethnically diverse populations.[20] The QDScore explained significantly more of the variation and had improved discrimination compared with the Cambridge risk score. Overall, almost half (15 545/32 450; 47.9%) of cases of diabetes occurred in the top 10th of the distribution and almost 70% (22 476/32 450) occurred in the top fifth based on the diabetes clinical risk score. This compared with 27.3% and 50% for the top 10th and fifth reported in the Cambridge risk score paper.[16] We cannot determine the calibration of the Cambridge risk score, as it does not give a measure of absolute risk over a given time period.

Our validation has some limitations. Although our validation cohort consisted of separate practices and patients, the practices used the same clinical computer system (EMIS) and so there may be a degree of overoptimism. Future studies could test the performance of the QDScore in other databases based on practices using a different clinical computer system or in cohorts in which formal diagnostic testing may be possible.

We did not do comparisons with other prospective studies that have developed a risk prediction score for which laboratory tests are needed (such as measurement of high density lipoprotein cholesterol,[17 49] triglycerides,[17 49] or fasting glucose[17]) or that have included variables which are difficult to measure consistently and reliably such as waist circumference and which, unlike body mass index, are not routinely recorded in general practice.[18 49] Other diabetes scores have been developed within specific ethnic groups (for example, Mexican Americans,[48] Japanese Americans[28]), but we have too few patients in the UK in these ethnic groups to allow a meaningful comparison to be made within this analysis. Nonetheless, our receiver operator curve statistic of 0.85 for women and 0.83 for men is substantially higher than those in many studies, which have reported values ranging between 0.71 and 0.80[16 27 28 49 60]; it is very comparable to the three studies reporting the highest receiver operator curve statistics, with values of 0.85 and 0.86.[17 18 48] Lastly, although data on fasting and random glucose are recorded to some extent within primary care electronic health

records,[25 61] we did not think that these were suitable for use in a prediction score, as they are the basis for making diagnoses of diabetes in this context rather than being recorded in a representative sample of patients at baseline. In addition, we were interested to develop a score that did not require laboratory measurements.

Conclusions

Simple risk algorithms have performed well in comparison with more complex clinical evaluations in studies of diabetes and cardiovascular disease.[17 47] This algorithm to predict risk of type 2 diabetes has the unique advantage of including both ethnicity and social deprivation, can be derived without laboratory measurements, and thus is suitable for use both in clinical settings and for self assessment. The QDScore could be used to identify patients at high risk of diabetes who might benefit from interventions to reduce their risk.

We acknowledge the contribution of Egton Medical Information System (EMIS) and practices using EMIS and contributing to the QResearch database.

Contributors: JH-C initiated and designed the study, obtained approvals, prepared the data, did the analysis and interpretation, and wrote the first draft of the paper. CC contributed to the development of the protocol, to the design, analysis, and interpretation, and to drafting the paper; she also did some of the primary analyses with JH-C. JR, PB, and AS contributed to the protocol, interpretation, and drafting the article. All authors approved the final draft. JH-C is the guarantor.

Funding: This study received no external funding. The authors did the work either in their personal time or during the course of their normal employment. The corresponding author (JH-C) and CC had access to all the data in the study, and all authors agreed and share responsibility for the decision to submit for publication.

Competing interests: JH-C is co-director of QResearch, a not for profit organisation, which is a joint partnership between the University of Nottingham and EMIS. JH-C is also director of ClinRisk Ltd, which produces software to ensure the reliable and updatable implementation of clinical risk algorithms within clinical computer systems to help to improve patients' care. EMIS is the leading supplier of information technology for 60% of UK general practices and may implement the QDScore within its clinical computer system. AS chairs the Equality and Diversity Forum of the National Clinical Assessment Service and is co-investigator on an MRC/NPRI funded randomised controlled trial aiming to prevent onset of type 2 diabetes in South Asians in the UK; he is also a co-investigator on the MRC Edinburgh Translational Medicine Methodology Hub. QResearch does analyses for the Department of Health and other government organisations. All research using QResearch is peer reviewed and published. This work and any views expressed within it are solely those of the co-authors and not of any affiliated bodies or organisations.

Ethical approval: The proposal was approved by the Trent Multi Centre Research Ethics Committee.

1 Hux JE, Bica A, Flintoft V, Ivis F. Population based estimates of the incidence and the prevalence of diabetes in Ontario. *Diabetes* 2000;49(suppl 1):A388.
2 Bagust A, Hopkinson P, Maslove L, Currie C. The projected health care burden of type 2 diabetes in the UK from 2000 to 2060. *Diabet Med* 2002;19:1-5.
3 Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. *N Engl J Med* 2007;357:370-9.
4 Kopelman PG. Obesity as a medical problem. *Nature* 2000;404:635-43.
5 Hossain P, Kawar B, El Nahas M. Obesity and diabetes in the developing world—a growing challenge. *N Engl J Med* 2007;356:213-5.
6 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.
7 Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 2002;346:393-403.
8 Tuomilehto J, Lindstrom J, Eriksson JG, Valle TT, Hamalainen H, Ilanne-Parikka P, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N Engl J Med* 2001;344:1343-50.
9 Chiasson J-L, Josse RG, Gomis R, Hanefeld M, Karasik A, Laakso M. Acarbose for prevention of type 2 diabetes mellitus: the STOP-NIDDM randomised trial. *Lancet* 2002;359:2072.
10 Ramachandran A, Snehlatha C, Mary S, Mukesh B, Bhaskar A, Vijay V. The Indian diabetes prevention programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1). *Diabetologia* 2006;49:289-97.
11 Gillies CL, Lambert PC, Abrams KR, Sutton AJ, Cooper NJ, Hsu RT, et al. Different strategies for screening and prevention of type 2 diabetes in adults: cost effectiveness analysis. *BMJ* 2008;336:1180-5.
12 Waugh N, Scotland G, McNamee P, Gillett M, Brennan A, Goyder E, et al. Screening for type 2 diabetes: literature review and economic modelling. *Health Technol Assess* 2007;11(17).
13 Weber MB, Narayan KMV. Diabetes prevention should be a public-health priority. *Lancet* 2008;371:473-4.
14 Gaziano TA, Galea G, Reddy KS. Scaling up interventions for chronic disease prevention: the evidence. *Lancet* 2007;370:1939-46.
15 Harris M. Undiagnosed NIDDM: clinical and public health issues. *Diabetes Care* 1993;16:642-52.
16 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.
17 Wilson PWF, 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.
18 Lindstrom J, Tuomilehto J. The diabetes risk score: a practical score to predict risk of type two diabetes. *Diabetes Care* 2003;26:725-31.
19 Glumer C, Carstensen B, Sandbaek A, Lauritzen T, Jorgensen T, Borh-Johnsend K. A Danish diabetes risk score for targeted screening: the inter99 study. *Diabetes Care* 2004;27:727-33.
20 Spijkerman A, Yuyun M, Griffin S, Dekker J, Nijpels G, Wareham N. The performance of a risk score as a screening test for undiagnosed hyperglycaemia in ethnic minority groups: data from the 1999 health survey for England. *Diabetes Care* 2007;27:116-22.
21 Smith JP. Economics of health and mortality special feature: nature and causes of trends in male diabetes prevalence, undiagnosed diabetes, and the socioeconomic status health gradient. *Proc Natl Acad Sci* 2007;104:13225-31.
22 Simmons D, Williams DRR, Powell MJ. Prevalence of diabetics in a predominantly Asian community: preliminary findings of the Coventry diabetes study. *BMJ* 1989;298:18-21.
23 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart* 2008;94:34-9.
24 Willi C, Bodenmann P, Ghali WA, Faris PD, Cornuz J. Active smoking and the risk of type 2 diabetes: a systematic review and meta-analysis. *JAMA* 2007;298:2654-64.
25 Hippisley-Cox J, Pringle M. Prevalence, care and outcomes for patients with diet controlled diabetes in general practice: cross sectional survey. *Lancet* 2004;364:423-28.
26 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;24:830-35.
27 Kanaya AM, Fyr CLW, de Rekeneire N, Shorr RI, Schwartz AV, Goodpaster BH, et al. Predicting the development of diabetes in older adults: the derivation and validation of a prediction rule. *Diabetes Care* 2005;28:404-8.
28 McNeely MJ, Boyko EJ, Leonetti DL, Kahn SE, Fujimoto WY. Comparison of a clinical model, the oral glucose tolerance test, and fasting glucose for prediction of type 2 diabetes risk in Japanese Americans. *Diabetes Care* 2003;26:758-63.
29 Conen D, Ridker PM, Mora S, Buring JE, Glynn RJ. Blood pressure and risk of developing type 2 diabetes mellitus: the women's health study. *Eur Heart J* 2007;28:2937-43.
30 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336:1475-82.
31 Weakliem DL. A critique of the bayesian information criterion for model selection. *Sociol Methods Res* 1999;27:359-97.
32 Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* 1999;28:964-74.
33 Gray A, Clarke P, Farmer A, Holman R, United Kingdom Prospective Diabetes Study Group. Implementing intensive control of blood glucose concentration and blood pressure in type 2 diabetes in England: cost analysis (UKPDS 63). *BMJ* 2002;325:860.
34 Royston P. Multiple imputation of missing values. *Stata J* 2004;4:227-41.
35 Schafer J, Graham J. Missing data: our view of the state of the art. *Psychol Methods* 2002;7:147-77.

36 Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. *J Epidemiol Community Health* 2007;60:979.
37 Moons KGM, Donders RART, Stijnen T, Harrell FJ. Using the outcome for imputation of missing predictor values was preferred. *J Epidemiol Community Health* 2006;59:1092.
38 Clark T, Altman D. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *J Epidemiol Community Health* 2003;56:28-37.
39 Royston P. Multiple imputation of missing values: update of ice. *Stata J* 2005;5:527-36.
40 Gail M, Pfeiffer R. On evaluating models of absolute risk. *Biostatistics* 2005;6:227-39.
41 Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999;18:2529-45.
42 Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;23:723-48.
43 Royston P. Explained variation for survival models. *Stata J* 2006;6:1-14.
44 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007;335:136.
45 Hoerger TJ, Harris R, Hicks KA, Donahue K, Sorensen S, Engelgau M. Screening for type 2 diabetes mellitus: a cost-effectiveness analysis. *Ann Intern Med* 2004;140:689-99.
46 Kannel WB, McGee DL. Diabetes and cardiovascular risk factors: the Framingham study. *Circulation* 1979;59:8-13.
47 Gaziano TA, Young CR, Fitzmaurice G, Atwood S, Gaziano JM. Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the NHANES I follow-up study cohort. *Lancet* 2008;371:923-31.
48 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.
49 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.
50 Rimm EB, Manson JE, Stampfer MJ, Colditz GA, Willett WC, Rosner B, et al. Cigarette smoking and the risk of diabetes in women. *Am J Public Health* 1993;83:211-4.
51 Wild S, MacLeod F, McKnight J, Watt G, MacKenzie C, Ford I, et al. Impact of deprivation on cardiovascular risk factors in people with diabetes: an observational study. *Diabet Med* 2008;25:194-9.
52 Doran T, Fullwood C, Gravelle H, Reeves D, Kontopantelis E, Hiroeh U, et al. Pay-for-performance programs in family practices in the United Kingdom. *N Engl J Med* 2006;355:375-84.
53 *Health survey for England: prevalence of undiagnosed diabetes by age and sex, adults aged 35 and over, 2003, England*. London: British Heart Foundation, 2003.
54 *Health survey for England: the health of minority ethnic groups*. Leeds: The Information Centre, 2004:435.
55 Ginde AA, Cagliero E, Nathan DM, Camargo CA Jr. Value of risk stratification to increase the predictive validity of HbA1c in screening for undiagnosed diabetes in the US population. *J Gen Intern Med* 2008;23:1346-53.
56 Kumari M, Head J, Marmot M. Prospective study of social and other risk factors for incidence of type 2 diabetes in the Whitehall II study. *Arch Intern Med* 2004;164:1873-80.
57 Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;342:1878-86.
58 Townsend P. Deprivation. *J Soc Policy* 1987;16:125-46.
59 Department of Health. *Health survey for England 2004*. DH: London, 2004.
60 Rahman M, Simmons R, Harding A, Wareham N, Griffin S. A simple risk score identifies individuals at high risk of developing type 2 diabetes: a prospective cohort study. *Fam Pract* 2008;25:191-6.
61 Holt TA, Stables D, Hippisley-Cox J, O'Hanlon S, Majeed A. Identifying undiagnosed diabetes: cross-sectional survey of 3.6 million patients' electronic records. *Br J Gen Pract* 2008;58:192-6.

Accepted: 19 January 2009