A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT)

Laura C Rosella,[1,2] Douglas G Manuel,[1,2,3,4] Charles Burchill,[5] Thérèse A Stukel,[1,6] for the PHIAT-DM team

ABSTRACT

Background National estimates of the upcoming diabetes epidemic are needed to understand the distribution of diabetes risk in the population and to inform health policy.

Objective To create and validate a population-based risk prediction tool for incident diabetes using commonly collected national survey data.

Methods With the use of a cohort design that links baseline risk factors to a validated population-based diabetes registry, a model (Diabetes Population Risk Tool (DPoRT)) was developed to predict 9-year risk for diabetes. The probability of developing diabetes was modelled using sex-specific Weibull survival functions for people >20 years of age without diabetes (N=19 861). The model was validated in two external cohorts in Ontario (N=26 465) and Manitoba (N=9899). Predictive accuracy and model performance were assessed by comparing observed diabetes rates with predicted estimates. Discrimination and calibration were measured using a C statistic and Hosmer–Lemeshow χ^2 statistic ($\chi 2_{H-L}$).

Results Predictive factors included were body mass index, age, ethnicity, hypertension, immigrant status, smoking, education status and heart disease. DPoRT showed good discrimination (C=0.77–0.80) and calibration (χ^2_{H-L} <20) in both external validation cohorts.

Conclusions This algorithm can be used to estimate diabetes incidence and quantify the effect of interventions using routinely collected survey data.

► Additional appendix are published online only. To view these files please visit the journal online (http://jech.bmj.com).

[1]Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada
[2]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[3]Ottawa Hospital Research Institute, Ottawa, Ontario, Canada
[4]Statistics Canada, Ottawa, Ontario, Canada
[5]University of Manitoba, Winnipeg, Manitoba, Canada
[6]Department of Health Policy, Management, and Evaluation, University of Toronto, Toronto, Ontario, Canada

Correspondence to
Dr Laura Rosella, G106, 2075 Institute for Clinical Evaluative Sciences, Bayview Avenue, Toronto, Ontario M4N 3M5, Canada;
laura.rosella@ices.on.ca

PHIAT-DM team members who contributed to the conception and draft of DPoRT are: Les Roos, Cam Mustard, Geoff Anderson, Jan Hux, Lisa Lix, Gillian Booth, Bernard Choi and Sarah Maaten

Accepted 2 February 2010
Published Online First 1 June 2010

INTRODUCTION

In medicine, prediction tools are used to calculate risk, defined as the probability of developing a disease or state in a given time period. Within the clinical setting, predictive tools such as the Framingham Heart Score[1] have contributed important advances in individual patient treatment and disease prevention.[2] Similarly, applying predictive risk tools to populations can provide insight into the influence of risk factors on the future burden of disease in an entire region or nation and the value of interventions at the population level.

Global estimates place the number of people with diabetes at approximately 200 million, and increasing rapidly.[3] There is a growing concern that these trends may slow or even reverse life expectancy gains in the USA and other developed countries.[4] Planning for healthcare and public health resources can be informed by robust prediction tools. Estimates of future diabetes incidence will alert policy makers, planners and physicians to the extent and urgency of the diabetes epidemic. In addition, a population prediction tool for diabetes can identify the optimal target groups for new intervention strategies, and determine how extensive a strategy must be to achieve the desired reduction in new cases. This insight can improve the effectiveness and efficiency of prevention strategies.

Clinical risk algorithms have been applied at the population level for other diseases,[5] but with considerable challenges. Clinical risk tools usually require clinical data that are rarely available at the population level. For diabetes, several clinical risk prediction tools exist, but they require clinical data that are collected infrequently or not at all at the population level, such as fasting blood sugar,[6–8] or require detailed information, such as diabetes family history.[9 10] In addition, some apply only to specific subgroups of the population, such as specific age ranges, or only to those with comorbid conditions.[11–13] For a population algorithm, the input variables should be representative of the entire population (ideally population-based), meaningful for health policy decision makers, available to a wide audience, and regularly collected so that estimates can be updated frequently. The creation and application of a population-based risk algorithm for diabetes is feasible because the risk factors for diabetes are well known and measured through self-reported questionnaires in population health surveys.

The objective of this study was to create a risk algorithm for diabetes incidence that can be applied at the level of populations using widely available public data. The Diabetes Population Risk Tool (DPoRT) was created and validated by individually linking three different provincial population health surveys to population-based registries of physician-diagnosed diabetes.

METHODS

DPoRT derivation cohort

The cohort was derived from 23 403 Ontario residents of the 1996/7 National Population Health Survey (NPHS-ON) conducted by Statistics Canada (83% response rate)[14] who were linkable to health administrative databases. Households were selected through stratified multilevel cluster sampling of residences using provinces and/or local planning regions as the primary sampling unit. The sample is proportionally representative of provinces according to the size of their populations. Excluded from the

Figure 1 Example use of the Diabetes Population Risk Tool to predict the 9-year risk of diabetes for a specific high-risk man.

Profile: Male; 55 years old; BMI = 29 kg/m², hypertension, white, smoker, heart disease, hypertension, and graduated secondary school

$$\text{Predicted risk (P)} = 1 - \exp(-\exp^{m})$$

$$m = \frac{\log(follow - uptime) - \mu}{scale}$$

$$\mu = 10.5971 - 0.2624 \times hypertension - 0.6316 \times non - white\ ethnicity - 0.5355 \times heart\ disease - 0.1765 \times smoker + 0.2344 \times secondary\ school\ education + 0.0000 \times BMI < 23 \times Age < 45 - 1.2378 \times 23 \le BMI < 25\ \&\ Age < 45 - 1.5490 \times 25 \le BMI < 30 \& Age < 45 - 2.5437 \times 30 \le BMI < 35 \times Age < 45 - 3.4717 \times BMI \ge 35 \times Age < 45 - 1.9794 \times BMI < 23\ \&\ Age \ge 45 - 2.4426 \times 23 \le BMI < 25\ \& Age \ge 45 - 2.8488 \times 25 \le BMI < 30\ \&\ Age \ge 45 - 3.3179 \times 30 \le BMI < 35\ \&\ Age \ge 45 - 3.5857 \times BMI \ge 35\ \&\ Age \ge 45$$

$$\mu = 10.5971 - 0.2624 \times (1 - 0.1084) - 0.6316 \times (0 - 0.1083) - 0.5355 \times (1 - 0.0519) - 0.1765 \times (1 - 0.2939) + 0.2344 \times (1 - 0.6289) - 1.2378 \times (0 - 0.1290) - 1.5490 \times (0 - 0.2207) - 2.5437 \times (0 - 0.0558) - 3.4717 \times (0 - 0.0120) - 1.9794 \times (0 - 0.068159) - 2.4426 \times (0 - 0.0860) - 2.8488 \times (1 - 0.2189925) - 3.3179 \times (0 - 0.0572891) - 3.5857 \times (0 - 0.0120)$$

$\mu = 8.9246$

$$m = \frac{\log(365.25 \times 9) - 8.9246}{0.8049}$$

$m = -1.0272$

9-year predicted risk for developing diabetes is:

$P = 1\text{-}\exp(\text{-}\exp^{\text{-}1.0272})$
$P = 30.09239$
or 30.09%

sampling frame were people living on Indian Reserves and Crown Lands, institutional residents, full-time members of the Canadian Forces, and residents of certain remote regions. The survey was conducted by telephone, and responses were self-reported. People under the age of 20 (n=2407), those with diabetes at baseline (n=894), women who were pregnant at baseline (n=241) (because baseline body mass index (BMI) could not be accurately ascertained), and men with missing baseline BMI (n=66) were excluded, resulting in 9177 male and 10 618 female subjects. Respondents were individually linked to a chart-validated population-based registry of physician-diagnosed diabetes. For further details of the breakdown of the derivation cohort, see figure 1 of the supplementary online appendix.

DPoRT validation cohorts

The first validation cohort was the Manitoba respondents of the 1996/7 NPHS (NPHS-MB) (N=10 118). The second validation cohort was from the Ontario portion of the 2000/1 Canadian Community Health Survey (CCHS-ON, Cycle 1.1, N=37 473, 81% response rate) administered by Statistics Canada. The target population of the CCHS is the same as that of the NPHS and provides data representative of 98% of the Canadian population[15 16] The same exclusion criteria were applied to both validation cohorts, and, after exclusions, there were 9899 in NPHS-MB and 26 465 in CCHS-ON. The NPHS-MB cohort had a 9-year follow-up (1996–2005), and CCHS-ON had a 5-year follow-up (2000–2005); therefore DpoRT-predicted risks were generated accordingly.

Identifying respondents who develop diabetes

Survey data were linked to provincial healthcare databases that include all people covered under the government-funded universal health insurance plan. The diabetes status of respondents in Ontario was established by linking people to the Ontario Diabetes Database (ODD), which contains all patients with physician-diagnosed diabetes identified since 1991. A patient is said to have physician-diagnosed diabetes if he or she meets at least one of the following criteria: (*a*) a hospital admission with a diabetes diagnosis (International Classification of Diseases Clinical Modification code 250 (ICD9-CM) before 2002 or ICD-10 code E10–E14 after 2002; (*b*) a physician services claim with a diabetes diagnosis (code 250) followed within 2 years by either a physician services claim or a hospital admission with a diabetes diagnosis. A hospital record with a diagnosis of pregnancy care or delivery close to a diabetic record (ie, a gestational admission date between 90 days before and 120 days after the diabetes record date) was considered to relate to a diagnosis of gestational diabetes and therefore the data were excluded. The ODD has been validated against primary care health records and demonstrated to be accurate for determining incidence and prevalence of diabetes (sensitivity 86%, specificity 97%).[17 18] Information on vital statistics and eligibility for healthcare coverage was captured from the Registered Persons Data Base. The ODD algorithm is applied nationally using provincial administrative registries (known as the National Diabetes Surveillance System) and has been successfully validated in several Canadian provinces.[19]

Variables

To ensure that DPoRT would be applicable across different populations, variables considered had to be based on established evidence, easily captured using population surveys, and captured in a consistent manner across surveys and populations. Variables included were age, height and weight, chronic conditions diagnosed by a health professional, ethnicity, immigration status, smoking status, educational achievement, household income, alcohol consumption and physical activity (based on metabolic equivalents). BMI in kg/m^2 was used as an indicator of obesity.[20] All variables were kept in the form released in the public use data file.

Statistical analysis

The probability of physician-diagnosed diabetes was assessed from the interview date until censoring for death or end of follow-up. The final model was fitted using a Weibull accelerated

Table 1 Baseline characteristics of development and validation cohorts

	Men			Women		
	Development cohort†	Validation cohorts‡		Development cohort	Validation cohorts	
Risk factor	Ontario NPHS (N=9177)	Manitoba NPHS (N=4670)	Ontario CCHS (N=12020)	Women NPHS (N=10618)	Manitoba NPHS (N=5229)	Ontario CCHS (N=14445)
Mean/median BMI (kg/m^2)	26.10/25.70	26.86/26.31	26.12/25.62	24.47/23.50	25.43/24.59	24.98/24.03
Mean age (years)	44	44	44	46	47	46
Age <45 (%)	54.80	55.67	55.85	51.68	52.71	51.59
Age 45–64 (%)	30.78	29.71	31.00	29.92	27.79	31.51
Age ≥65 (%)	14.42	14.63	13.15	18.39	19.51	16.90
BMI <23 (%)	19.48	17.79	22.23	40.39	35.89	39.29
BMI 23–24 (%)	22.11	20.34	21.51	19.01	16.65	17.79
BMI 25–29 (%)	43.97	44.34	40.03	24.36	28.51	27.19
BMI 30–34 (%)	11.31	14.40	12.74	8.50	10.55	9.47
BMI ≥35 (%)	2.40	2.63	3.05	2.77	3.15	4.11
BMI missing (%)	0.73	0.51	0.44	4.98	5.26	2.14
Non-white (%)	11.51	10.42	16.68	10.41	10.51	16.76
Hypertension (%)	10.23	10.16	12.50	12.32	13.22	14.94
Current smoker (%)	29.67	30.76	24.78	24.48	24.40	18.97
Physical activity (kcal/day)	1.86/1.20	1.79/1.10	1.97/1.30	1.62/1.10	1.44/1.00	1.63/1.10
Heart disease (%)	4.97	4.43	5.19	4.16	4.62	5.24
Graduated post secondary school (%)	81.12	73.28	82.11	81.86	73.81	81.09
Number incident diabetes (unweighted)	718	272	559	692	258	558
% developing diabetes in 9 years	7.78	7.22		6.13	4.75	
Age standardised* % developing diabetes in 9 years	6.67	6.55		5.59	4.27	
% developing diabetes in 5 years	4.26		4.60	3.23		3.69
Age standardised* % developing diabetes in 5 years	3.59		3.95	2.81		3.35

Categorical variables are represented as a proportion (%), and continuous variables are represented as a mean/median.
*Standardised to the 1991 Canadian population.
†The development cohort refers to the cohort where the Diabetes Population Risk Tool (DPoRT) was created: the Ontario 1996/7 National Population Health Survey (NPHS) linked to diabetes status.
‡The validation cohorts refer to the populations that DPoRT was validated on: the Manitoba 1996/7 NPHS and the Ontario 2000/1 Canadian Community Health Survey (CCHS) both linked with diabetes status.
BMI, body mass index.

failure time model which is a time-to-event model that predicts probability of diabetes and includes time in its equation, allowing the user to predict diabetes probability for a range of follow-up periods. Diabetes functions were derived separately for men and women. Variables were added to the model in a nested fashion based on clinical importance, and the marginal statistical and predictive significance was evaluated, controlling for variables already in the model. Predicted probability for each person was calculated by multiplying their risk factor values by the corresponding regression coefficients and summing the products.[21] An example of this calculation is shown in figure 1. The functional form of the model was assessed using likelihood ratio tests to compare nested parametric models.[22] A plot of log(−logS(t)) versus log(t) (where S(t) = the probability of being diabetes free beyond time t) and Cox–Snell residual plots were produced to assess the Weibull distribution. Two indices of model performance were examined: discrimination and calibration. Model discrimination is the ability to correctly classify those with and without the disease based on predicted risk—that is, to rank those who will and will not develop diabetes. Discrimination is measured using a C statistic, which is analogous to the area under the receiver operating characteristic curve, a plot of sensitivity versus (1−specificity) for a binary outcome at various thresholds.[23] This study uses a C statistic modified for survival data developed by Pencina and D'Agostino.[24] Calibration describes the accuracy of a prediction model—specifically, the extent of agreement between predicted and observed outcomes. It is measured using the Hosmer and Lemeshow statistic (H–L test), a χ^2 test based on grouping observations into deciles of predicted risk and testing associations with observed outcomes.[25] In our study, it was calculated by comparing observed diabetes rates and DpoRT-predicted diabetes probabilities using a modified version of the H–L χ^2 statistic for time-to-event data.[26 27] To mark sufficient calibration, χ^2=20 was used as a cut-off ($p<0.01$), consistent with the method of D'Agostino *et al*[26] in validating the Framingham algorithms. In addition, discrimination and calibration were computed using the coefficients generated from the validation cohort and labelled 'own cohort'. This was done to assess if the coefficients generated from the validation cohort produced significantly different predictive accuracy from DPoRT. In addition, graphical representations of predicted and observed rates were produced. Recalibration was achieved by substituting the mean values from the validation cohort to define all variables. Because of systematic case ascertainment differences between provinces, a further adjustment was applied to predicted rates outside of Ontario. To demonstrate the applicability of this tool, the sex-specific DPoRT models were applied to the most recently released national survey (CCHS 2005) to provide estimates of diabetes incidence up to 2014. Predicted diabetes cases and incidence rates were generated overall and by age, BMI, ethnicity and education.

All estimates incorporated bootstrap replicate survey weights to accurately reflect the demographics of the population and account for the survey sampling design based on selection probabilities and post-stratification adjustments. Variance estimates were calculated using bootstrap survey weights.[28 29] All statistics were computed using SAS statistical software (version 9.1).

Table 2 Diabetes Population Risk Tool functions for predicting 9-year risk of physician-diagnosed diabetes for men

Risk factor	Value
Intercept	10.5971
Hypertension	
No	0.00
Yes	−0.2624
Non-white ethnicity	
No	0.00
Yes	−0.6316
Heart disease	
No	0.00
Yes	−0.5355
Current smoker	
No	0.00
Yes	−0.1765
Education	
< Post-secondary	0.00
Post-secondary or higher	0.2344
BMI/age category	
BMI <23/age <45	0.00
BMI 23–24/age <45	−1.2378
BMI 25–29/age <45	−1.5490
BMI 30–34/age <45	−2.5437
BMI ≥35/age <45	−3.4717
BMI <23/age ≥45	−1.9794
BMI 23–24/age ≥45	−2.4426
BMI 25–29/age ≥45	−2.8488
BMI 30–34/age ≥45	−3.3179
BMI ≥35/age ≥45	−3.5857
Scale	0.8049

BMI, body mass index (m/kg^2).

Table 3 Diabetes Population Risk Tool functions for predicting 9-year risk of physician-diagnosed diabetes for women

Risk factor	Value
Intercept	10.5474
Hypertension	
Yes	0.00
No	−0.2865
Non-white ethnicity	
Yes	0.00
No	−0.4309
Immigrant status	
Yes	0.00
No	−0.2930
Education	
< Post-secondary	0.00
Post-secondary or higher	0.2042
BMI/age category	
BMI <23/age <45	0.00
BMI 23–24/age <45	−0.5432
BMI 25–29/age <45	−0.8453
BMI 30–34/age <45	−1.4104
BMI ≥35/age <45	−2.0483
BMI missing/age <45	−1.1328
BMI<23/age 45–64	0.0711
BMI 23–24/age 45–64	−0.7011
BMI 25–29/age 45–64	−1.4167
BMI 30–34/age 45–64	−2.2150
BMI ≥35/age 45–64	−2.2695
BMI missing/age 45–64	−1.7260
BMI <23/age ≥65	−1.0823
BMI 23–24/age ≥65	−1.1419
BMI 25–29/age ≥65	−1.5999
BMI 30–34/age ≥65	−1.9254
BMI ≥35/age ≥65	−2.1959
BMI missing/age ≥65	−1.8284
Scale	0.7814

BMI, body mass index (m/kg^2).

RESULTS

All population characteristics in the derivation cohort and two validation cohorts are shown in table 1. In the development cohort, 7.78% and 6.13% of men and women, respectively, developed diabetes during the follow-up period. The derivation cohort differed from the validation cohorts in risk factor and sociodemographic composition, particularly ethnic composition (table 1).

Diabetes risk was strongly related to BMI and age. BMI was considered in both its continuous and categorical form; however, the best goodness-of-fit and calibration were achieved by categorising BMI and including its interactions with age. This categorisation was least likely to cause over-fitting when applied to external data while maintaining discrimination. Including only age and BMI achieved a moderate degree of discrimination (C statistic=0.70). Non-white ethnicity, hypertension and less than post-secondary education were also important factors associated with an increased risk of diabetes. For men, smoking and heart disease were important independent risk factors found to improve model characteristics; for women, immigrant status improved the model. The following variables were excluded because they did not improve the model or worsened predictive accuracy: income, physical activity and alcohol consumption. The DPoRT model for predicting diabetes risk is shown in table 2 for men and table 3 for women. Figure 1 demonstrates how the risk coefficients in DPoRT were used to calculate risk using a man with numerous risk factors as an example.

In the CCHS validation cohort, the DPoRT 5-year predicted (and observed) diabetes incidence rates were 4.2% (4.6%) for men and 3.4% (3.7%) for women. In the NPHS-MB validation cohort, DPoRT 9-year predicted (observed) diabetes incidence rates were 7.0% (7.1%) for men and 5.1% (4.7%) for women. Overall predicted diabetes rates differed from observed rates by ≤0.4% in both validation cohorts (figure 2 of the supplementary online appendix).

Observed and predicted risks closely agreed overall and across levels of diabetes risk. R^2 between observed diabetes rates and DpoRT-predicted probabilities across quantiles of risk exceeded 98%. C statistics when DPoRT was applied to the validation cohorts were high (0.77–0.80) and were not appreciably lower than those generated from the 'own cohort' models (table 4). As shown graphically in figure 2, for men and women, observed and predicted rates of diabetes did not substantially differ across quantiles of risk in both validation cohorts ($\chi_{H-L}<20$).

To demonstrate the applicability of DPoRT to recent Canadian data, the predicted risk and number of new cases in the next 9 years using the 2005 data are shown in table 5. Approximately 1.7 million new diabetes cases are predicted for the subsequent 9 years with significant variability by age, ethnicity, education and levels of obesity (table 5). Risk increases sharply with increasing levels of obesity, and decreases with increasing level of educational achievement.

DISCUSSION

This study shows that diabetes risk can be accurately predicted at the population level using self-reported measures available in

Table 4 C statistics with 95% CIs and calibration $\chi 2$ statistics for Diabetes Population Risk Tool (DpoRT) and cohorts' own functions

	Men			Women		
	NPHS ON	CCHS ON	NPHS MB	NPHS ON	CCHS ON	NPHS MB
C statistic (95% CI)						
DPoRT	0.77 (0.76 to 0.79)	0.77 (0.76 to 0.79)	0.79 (0.77 to 0.82)	0.78 (0.76 to 0.79)	0.76 (0.74 to 0.77)	0.80 (0.77 to 0.82)
Own function*	0.80 (0.78 to 0.83)	0.78 (0.76 to 0.79)	0.80 (0.77 to 0.82)	0.80 (0.78 to 0.83)	0.77 (0.75 to 0.79)	0.80 (0.77 to 0.82)
Calibration $\chi 2$						
Uncalibrated DPoRT	4.33	13.23	136.13	5.22	24.84	35.07
Mean calibrated DPoRT‡	—	13.04	18.35	—	18.27	17.61
Own function	—	8.89	8.32	—	10.44	4.88

*Own function is the factors of the algorithm applied using coefficients derived from the validation cohort's own data.
‡Calibrated DPoRT is function adjusted using the validation cohort's own means for factors.
CCHS, Canadian Community Health Survey; MB, Manitoba; NPHS, National Population Health Survey; ON, Ontario.

population health surveys. In addition to being able to effectively rank-order subjects from low to high risk, DPoRT-predicted rates closely agreed with observed rates across levels of risk for both sexes in two external validation cohorts. DPoRT represents a novel approach that can be integrated into commonly collected population health survey data.

Similar to the clinical setting, where decisions are guided by estimation of baseline risk at the time of patient assessment, DPoRT allows estimation of baseline risk of diabetes using the current level of risk factors in the population. This approach has several advantages. Firstly, it has been widely accepted with demonstrated utility for clinical decision-making by providing an effective way to assess patient risk based on multiple risk factors. This risk assessment is then used to guide treatment or prevention recommendations. Secondly, there are well-described methods for developing clinical risk tools and assess their validity, providing a sound methodological basis to validate our approach.[26 30] Thirdly, the use of simple self-reported risk factor data allows DPoRT to be used in population settings where detailed clinical data are often unavailable. Further, the use of regularly collected population surveys allows estimates to be updated frequently. The most stringent test of predictive model

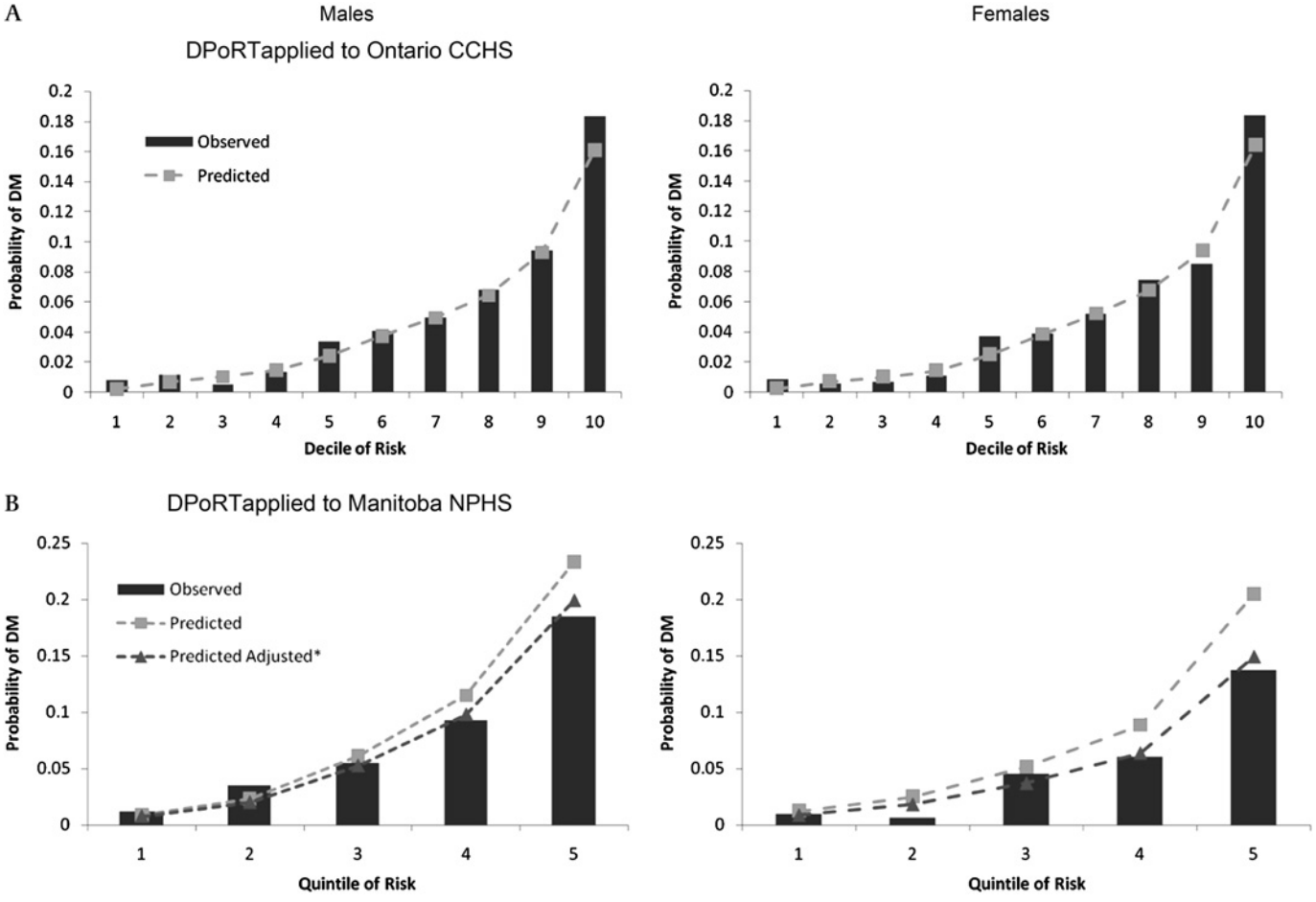

Figure 2 Predicted 10 versus observed incidence of diabetes for men and women in two validation datasets across deciles or quintiles of risk. The x axis refer to quantile (decile or quintile) of predicted Diabetes Population Risk Tool (DPoRT). The y axis refers to the observed (bars) and DPoRT-predicted (dotted line) probability of developing Diabetes Mellitus (DM) in a 5-year period for Ontario and a 9-year period for Manitoba. Observed diabetes rates are physician-diagnosed diabetes rates in the same time period. CCHS, Canadian Community Health Survey; DPoRT, Diabetes Population Risk Tool; NPHS, National Population Health Survey.

Table 5 Predicted 9-year diabetes risk in 2005 by subgroups in Canada from the Canadian Community Health Survey (CCHS)

	Men		Women	
Characteristic	9-year risk (%)	No of new cases	9-year risk (%)	No of new cases
Age group				
<35	3.4	104277	3.9	112821
35–54	8.5	386398	6.3	282964
55–75	14.4	334362	10.8	275086
>75	13.6	69917	11.9	100537
Body mass index				
<23	2.7	57028	2.7	102262
23–24	5.5	122731	4.4	79762
25–29	8.8	384631	9.2	254164
30–34	17.3	230934	17.6	180091
≥35	28.3	99629	22.6	100734
Ethnicity				
White	8.2	712985	7.4	611092
Non-white	10.7	181968	9.7	160316
Education level				
< Secondary	13.2	211818	12.1	209799
Secondary school graduation	9.5	152154	8.7	149895
Other post-secondary school	6.6	60718	6.1	49803
Graduated post-secondary school	7.4	470263	5.9	361911
Overall	8.6	894953	7.0	771408

accuracy is the application of the model to a different population.[31 32] This study shows that DPoRT is discriminating and accurate in two external populations that varied across geography and time.

Previous studies that estimate future diabetes burden have either extrapolated overall trends in diabetes prevalence or indirectly incorporated information on the influence of risk factors with various assumptions.[3 33–36] Studies of diabetes lifetime risk and life expectancy are not predictive; rather they describe diabetes from a life-course perspective using a period or stationary population approach.[36 37] Although these approaches are useful, they do not enable users to directly and quantitatively assess the impact of risk factors, such as BMI, on future diabetes cases. Complex modelling and simulation methods differ from the approach used in this study in that they use additional information on how populations and risk factors change over time.[38 39] A strength of simulation models is that they can combine different data sources.[40] However, these models often represent clinical or theoretical populations, making estimates difficult to validate in populations that are meaningful for population health planning.

Potentially important clinical values, such as fasting blood glucose, are excluded from DPoRT because they are not captured in population surveys. Although these variables may be clinically important, their use is not feasible for population risk assessment because they are not routinely collected in most populations. These omitted variables are unlikely to have a major effect on the performance characteristics of the model because of the clustering of risk factors, particularly when dealing with abnormalities of the metabolic system.[41–45]

A potential limitation of this study is that variables such as family history of diabetes or poor diet were not collected. These variables are also associated with the clustering of metabolic and other risk factors included in the algorithm. Using self-report measures is a limitation because these measures may be subject to reporting error. Validation studies have shown a strong correlation between measured height and weight; however, weight has been shown to be underestimated and height overestimated.[46 47] It is important to note that DPoRT is designed to be applied to self-reported data, and, unless surveys or reporting patterns change, this is unlikely to affect model performance. There is evidence that simple clinical risk tools, including those with self-report data, perform as well as complex models.[48 49]

To ensure that DPoRT can be applied in different populations, we used variables that remained stable over time, were unlikely to be subject to serious measurement error (such as alcohol and dietary habits), and are easily captured using survey data in different populations. For example, physical activity, shown to have a protective effect on diabetes risk, was removed from the model because of the inability to capture it in a reliable and reproducible manner across surveys, as well as its lack of improvement of model accuracy. Despite considerable variable constraints, DPoRT maintained good discrimination. Additional predictive variables need to have a high independent risk (OR ≥6.9) to result in significant improvements once a discrimination of 0.8 is already achieved.[50] This phenomenon was corroborated in this study, as a maximum level of discrimination was achieved using few variables.

DPoRT was developed in Canada and is most appropriate in the Canadian setting; however, like other risk tools, it may be transportable once validated and calibrated. The simplicity of the model and the fact that it was validated in two very different populations make its generalisability a genuine possibility. We recommend that, when this model is applied outside of Canada, it be validated to ensure accuracy.

The use of physician-diagnosed diabetes, as opposed to true diabetes status (diagnosed plus undiagnosed), is a limitation

What is already known on this subject

- Risk algorithms are often used in the clinical setting to guide clinical decision making; however, they have typically not been adapted for use at the population level.
- Clinical risk tools for diabetes require data that are rarely available at the population level.
- Previous studies that have estimated future diabetes burden have either extrapolated overall trends in diabetes prevalence or indirectly incorporated information on the influence of risk factors with various assumptions.

What this study adds

- This study develops and validates an algorithm for population-based prediction of diabetes (DPoRT), which differs from individual risk approaches.
- This algorithm can be used to estimate diabetes incidence in populations and quantify the effect of interventions using routinely collected survey data on risk factors.
- Population-based prediction models such as DPoRT can be used to improve population health planning, explore the effect of prevention strategies, and enhance our understanding of the distribution of diabetes in the population.

because the estimates may exclude people with diabetes who are not yet identified. This may reflect patients with less severe disease and/or poorer access to medical care. Physician-diagnosed diabetes is currently the most commonly used definition of diabetes at the level of populations. Although true prevalence estimates would be higher, advocates of the physician-diagnosed outcome argue that it is meaningful to people with recognised diabetes and to the treatment of patients in the healthcare system. In Canada, all residents are covered under a universal health insurance plan and thus are eligible for healthcare and access to a physician for diabetes testing. If diabetes testing/screening increases over time, predicted estimates may be lower than the observed estimates (under the assumption of increased case detection). DPoRT has been found to be accurate in different populations for different time periods; however, it could be adjusted to predict total diabetes cases using information on screening/testing in the population.

Curbing the diabetes epidemic has been identified by governments and health policy makers as a top priority for improving, and even maintaining, the health of their nations. Population-based prediction models such as DPoRT can be used to improve health planning, explore the effect of prevention strategies, and enhance understanding of distribution of the disease. This study shows that DPoRT accurately predicts diabetes incidence and is effective at predicting the population level of diabetes risk. This algorithm can be used by health planners to estimate diabetes incidence, to stratify the population by risk, and quantify the effect of interventions using routinely collected survey data. As the surveillance of risk factors and diabetes advances, DPoRT can be adapted to become more accurate, while maintaining its accessibility for decision makers.

Acknowledgements We thank Dr Michael J. Pencina, Boston University, for providing the SAS macro for calculating survival-based calibration and discrimination, Mr Refik Saskin, Institute for Clinical Evaluative Sciences, for assistance with data acquisition, and Amy Zierler for editing the manuscript.

Funding The study is funded by the Canadian Institutes of Health Research. DM holds a Chair in Applied Public Health from the Canadian Institute for Health Research and Public Health Agency of Canada. The views expressed here are those of the authors and not necessarily those of the funding agency. The funding agency had no role in the data collection or in the writing of this paper. The guarantors accept full responsibility for the conduct of the study, had access to the data, and controlled the decision to publish.

Competing interests None.

Ethical approval The study was approved by the Research Ethics Board of Sunnybrook Health Sciences Centre, Toronto, Ontario and at the University of Manitoba - Research Ethics Board in Winnipeg, Manitoba.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Anderson KM,** Wilson PWF, Odell PM, *et al*. An Updated Coronary Risk Profile - A Statement for Health-Professionals. *Circulation* 1991;**83**:356—62.
2. **Hippisley-Cox J,** Coupland C, Vinogradova Y, *et al*. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *Br Med J* 2007;**335**:136—41.
3. **Wild S,** Roglic G, Green A, *et al*. Global prevalence of diabetes - Estimates for the year 2000 and projections for 2030. *Diabetes Care* 2004;**27**:1047—53.
4. **Zimmet P,** Alberti KGMM, Shaw J. Global and societal implications of the diabetes epidemic. *Nature* 2001;**414**:782—7.
5. **Manuel DG,** Kwong K, Tanuseputro P, *et al*. Effectiveness and efficiency of different guidelines on statin treatment for preventing deaths from coronary heart disease: modelling study. *BMJ* 2005;**332**:1419—22.
6. **Eddy DM,** Schlessinger L. Validation of the archimedes diabetes model. *Diabetes Care* 2003;**26**:3102—10.
7. **Hanley AJG,** Williams K, Gonzalez C, *et al*. Prediction of type 2 diabetes using simple measures of insulin resistance - combined results from the San Antonio Heart Study, the Mexico city diabetes study, and the insulin resistance atherosclerosis study. *Diabetes* 2003;**52**:463—9.
8. **Ito C,** Maeda R, Nakamura K, *et al*. Prediction of diabetes mellitus (NIDDM). *Diabetes Res Clin Pract* 1996;**34** (Suppl):S7—11.
9. **Herman WH,** Smith PJ, Thompson TJ, *et al*. A new and simple questionnaire to identify people at increased risk for undiagnosed diabetes. *Diabetes Care* 1995;**18**:382—7.
10. **Lindstrom J,** Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care* 2007;**26**:725—31.
11. **Larsson H,** Lindgerde F, Berglund G, *et al*. Prediction of diabetes using ADA or WHO criteria in post-menopousal women: a 10-year follow-up study. *Diabetologia* 2000;**43**:279—88.
12. **Stern MP,** Williams K, Haffner SM. Identification of persons at high risk of type 2 diabetes mellitus: Do we need the oral glucose tolerance test. *Ann Intern Med* 2009;**136**:575—81.
13. **Wilson P,** Meigs JB, Sullivan LM, *et al*. Prediction of incident diabetes mellitus in middle-aged adults. *Arch Intern Med* 2007;**167**:1068—74.
14. **Statistics Canada.** *1996-97 NPHS public use microdata documentation*. Ottawa: Statistics Canada, 1999.
15. **Statistics Canada.** Canadian community health survey methodological overview. *Health Reports* 2002;**13**:9—14.
16. **Statistics Canada.** *Canadian community health survey 2000—2001. Statistics Canada*. ed. Ottawa: Statistics Canada, 2003. Ref Type: Data File.
17. **Hux JE,** Ivis F. Diabetes in Ontario. *Diabetes Care* 2005;**25**:512—16.
18. **Lipscombe LL,** Hux JE. Trends in diabetes prevalence, incidence, and mortality in Ontario, Canada 1995-2005: a population-based study. *Lancet* 2007;**369**:750—16.
19. **Health Canada.** *Responding to the challenge of diabetes in Canada*. Ottawa, ON, 2003.
20. **Statistics Canada.** *1996-7 National population health survey*. Ottawa: Derived Variable Specifications, 1999.
21. **Odell PM,** Anderson KM, Kannel WB. New models for predicting cardiovascular events. *J Clin Epidemiol* 1994;**47**:583—92.
22. **Farewell VT,** Prentice RL. A study of distributional shape in life testing. *Technometrics* 1977;**19**:69—75.
23. **Campbell G.** General methodology I: advances in statistic methodology for the evaluation of diagnostic and laboratory tests. *Stat Med* 2004;**13**:499—508.
24. **Pencina M,** D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* 2004;**23**:2109—23.
25. **Hosmer DW,** Hosmer T, Cessie LE, *et al*. A comparison of goodness-of-ft tests for the logistic regression model. *Stat Med* 1997;**16**:965—80.
26. **D'Agostino RB,** Grundy S, Sullivan LM, *et al*. Validation of the framingham coronary disease prediction scores. *JAMA* 2001;**286**:180—7.
27. **Nam BH.** *Discrimination and calibration in survival analysis: extension of the ROC curve for descrimination and chi-square test for calibration boston University* 2000.
28. **Yeo D,** Mantel H, Lui TP. Bootstrap variance estimation for the national population health survey. Baltimore: American Statistical Association, 1999.
29. **Kovacevic MS,** Mach L, Roberts G. Bootstrap variance estimation for predicted individual and population-average risks. *Proceedings of the Survey Research Methods Section*. Alexandria VA: American Statistical Association, 2008.
30. **D'Agostino RB,** Griffith JL, Schmidt CH, *et al*. *Measures for evaluating model performance. proceedings of the biometrics section*. Alexandria, VA: American Statistical Association, Biometrics Section, 1997:253—8.
31. **Altman DG,** Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;**19**:453—73.
32. **Harrell FE,** Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361—87.
33. **Boyle JP,** Honeycutt AA, Narayan KMV, *et al*. Projection of diabetes burden through 2050-Impact of changing demography and disease prevalence in the US. *Diabetes Care* 2001;**24**:1936—40.
34. **King H,** Aubert RE, Herman WH. Global burden of diabetes, 1995-2025-prevalence, numerical estimates, and projections. *Diabetes Care* 1998;**21**:1414—31.
35. **World Health Organization.** *Report of a WHO consultation on obesity, obesity: preventing and managing the global epidemic.* Geneva: World Health Organization, 1998.
36. **Narayan KMV,** Boyle JP, Thompson TJ, *et al*. Lifetime risk for diabetes mellitus in the United States. *JAMA* 2003;**290**:1884—90.
37. **Manuel D,** Schultz S. Health-related quality of life and health-adjusted life expectancy of people with diabetes in Ontario, Canada, 1996-1997. *Diabetes Care* 2004;**27**:407—14.
38. **Forouhi NG,** Merrick D, Goyder E, *et al*. Diabetes prevalence in England, 2001-estimates from an epidemiological model. *Diabet Med* 2006;**23**:189—97.
39. **Mainous AG,** Baker R, Koopman RJ, *et al*. Impact of the population at risk of diabetes on projections of diabetes burden in the United States: an epidemic on the way. *Diabetologia* 2007;**50**:934—40.
40. **Ford ES,** Ajani UA, Croft JB, *et al*. Explaining the decrease in US deaths from coronary disease, 1980—2000. *N Engl J Med* 2007;**356**:2388—98.
41. **Carmelli D,** Cardon LR, Fabsitz R. Clustering of hypertension, diabetes, and obesity in adult male twins - same genes or same environments. *Am J Hum Genet* 1994;**55**:566—73.
42. **DeFronzo RA,** Ferrannini E. Insulin resistance. A multifaceted syndrome responsible for NIDDM, obesity, hypertension, dyslipidemia, and atherosclerotic cardiovascular disease. *Diabetes Care* 1991;**41**:173—94.

43. **Lorenzo C,** Okoloise M, Williams K, *et al*. The metabolic syndrome as predictor of type 2 diabetes - The San Antonio heart study. *Diabetes Care* 2003;**26**: 3153—9.
44. **Meigs JB,** Dagostino RB, Wilson PWF, *et al*. Risk variable clustering in the insulin resistance syndrome - the framingham offspring study. *Diabetes* 1997;**46**:1594—600.
45. **Schmidt MI,** Watson RL, Duncan BB, *et al*. Clustering of dyslipidemia, hyperuricemia, diabetes, and hypertension and its association with fasting insulin and central and overall obesity in a general population. *Metab Clin Exp* 1996;**45**:699—706.
46. **Nawaz H,** Chan W, Abdulraham M, *et al*. Self-reported weight and height: implications for obesity research. *J Prev Med* 2001;**20**:294—8.
47. **Rowland M.** Self-reported height and weight. *Am J Clin Nutr* 2007;**52**:1125—33.
48. **Ambler G,** Brady AR, Royston P. Simplifying a prognostic model: a simulation model based on clinical data. *Stat Med* 2002;**21**:3803—22.
49. **Mainous AG,** Koopman RJ, *et al*. Coronary heart disease risk score based on pateint-reported information. *Am J Cardiol* 2007;**99**:1236—41.
50. **Pepe MS,** Janes H, Longton G, *et al*. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004;**159**:882—90.