BMJ Open Diabetes Research & Care

External validation and application of the Diabetes Population Risk Tool (DPoRT) for prediction of type 2 diabetes onset in the US population

Kathy Kornas,[1] Christopher Tait,[1] Ednah Negatu,[1] Laura C Rosella [1,2,3,4]

To cite: Kornas K, Tait C, Negatu E, *et al*. External validation and application of the Diabetes Population Risk Tool (DPoRT) for prediction of type 2 diabetes onset in the US population. *BMJ Open Diab Res Care* 2024;**12**:e003905. doi:10.1136/bmjdrc-2023-003905

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjdrc-2023-003905).

Received 10 November 2023
Accepted 20 February 2024

[1]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[2]ICES, Toronto, Ontario, Canada
[3]Institute for Better Health, Trillium Health Partners, Mississauga, Ontario, Canada
[4]Temerty Faculty of Medicine, Department of Laboratory Medicine and Pathobiology, Toronto, Ontario, Canada

Correspondence to
Dr Laura C Rosella;
laura.rosella@utoronto.ca

ABSTRACT

Introduction Characterizing diabetes risk in the population is important for population health assessment and diabetes prevention planning. We aimed to externally validate an existing 10-year population risk model for type 2 diabetes in the USA and model the population benefit of diabetes prevention approaches using population survey data.

Research design and methods The Diabetes Population Risk Tool (DPoRT), originally derived and validated in Canada, was applied to an external validation cohort of 23 477 adults from the 2009 National Health Interview Survey (NHIS). We assessed predictive performance for discrimination (C-statistic) and calibration plots against observed incident diabetes cases identified from the NHIS 2009–2018 cycles. We applied DPoRT to the 2018 NHIS cohort (n=21 187) to generate 10-year risk prediction estimates and characterize the preventive benefit of three diabetes prevention scenarios: (1) community-wide strategy; (2) high-risk strategy and (3) combined approach.

Results DPoRT demonstrated good discrimination (C-statistic=0.778 (males); 0.787 (females)) and good calibration across the range of risk. We predicted a baseline risk of 10.2% and 21 076 000 new cases of diabetes in the USA from 2018 to 2028. The community-wide strategy and high-risk strategy estimated diabetes risk reductions of 0.2% and 0.3%, respectively. The combined approach estimated a 0.4% risk reduction and 843 000 diabetes cases averted in 10 years.

Conclusions DPoRT has transportability for predicting population-level diabetes risk in the USA using routinely collected survey data. We demonstrate the model's applicability for population health assessment and diabetes prevention planning. Our modeling predicted that the combination of community-wide and targeted prevention approaches for those at highest risk are needed to reduce diabetes burden in the USA.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ The Diabetes Population Risk Tool (DPoRT) is a population-based risk algorithm that was originally validated in Canada to predict 10-year incidence of physician-diagnosed type 2 diabetes using routinely collected population survey data.

WHAT THIS STUDY ADDS

⇒ DPoRT accurately predicted diabetes risk in the US population and the model estimated that combining community-wide and high-risk prevention strategies would prevent the most diabetes cases in 10 years.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ DPoRT can reliably be used in the USA for population-based diabetes risk assessment and for informing population-level diabetes prevention.

INTRODUCTION

The prevalence of diabetes has increased over the last several decades in all regions of the world.[1 2] In the USA, 14% of the adult population was living with type 2 diabetes in 2017/18,[3] with trends showing increases in diabetes incidence between 1990 and 2017.[4] The urgent need for population-wide diabetes prevention is reflected in the Affordable Care Act's provisions, expanding access to clinical and community preventive services and the establishment of the National Diabetes Prevention Program.[5]

Characterizing diabetes risk in the population is important for informing the right mix of preventive strategies, which range from individualized interventions targeted to those with high risk (eg, pharmacotherapy) to community-wide public policy interventions to reduce key risk factors in the whole population, such as physical inactivity and obesity.[6 7] Population-level risk algorithms are specifically designed for estimating baseline risk for the whole population and are useful for modeling the population benefits from prevention strategies.[8] One such tool is the Diabetes Population Risk Tool (DPoRT), a multivariable risk algorithm that estimates 10-year diabetes risk in populations using self-reported risk factor information that is routinely collected in population health surveys.[7 9] DPoRT has demonstrated applicability for population health assessment and planning, with uptake by major public health

units and other health settings in Canada to understand how diabetes risk is distributed in local communities.[10]

Before applying a prediction model in a new geographic setting, it is essential to assess the accuracy of the model's predictions in the new population.[11] External validation of prediction models in a different population from which the model was originally derived is also important for demonstrating generalizability in new settings.[11] However, a systematic review found that only a small portion of existing prediction models were externally validated in independent datasets.[12] In addition, it is a more common practice to re-derive models, as opposed to validating existing prediction models for new settings, which results in redundant models with similar results and additional resources.[13] DPoRT was developed and externally validated in Canada because of the ability to probabilistically link population health data with administrative data for physician-diagnosed diabetes to allow for validation. Given that linkages of this nature are not readily available in all jurisdictions, a novel approach is necessary for model evaluation of this risk algorithm in new populations.

The present study focuses on examining the validity of DPoRT for the US population and describes a methodology that may inform model evaluation and updating of population risk prediction tools internationally. The objectives of this study were to externally validate DPoRT in a nationally representative cohort of the US population and to demonstrate the utility of DPoRT by modeling prevention strategies for type 2 diabetes in the USA.

RESEARCH DESIGN AND METHODS

Context and setting

We previously developed and validated DPoRT in the Canadian population.[7 9] DPoRT is a population-based risk tool that estimates the 10-year incidence of physician-diagnosed type 2 diabetes. DPoRT predicts the probability of developing diabetes using a sex-specific statistical model based on the Weibull survival distribution for people 20 years and older. The original risk algorithm was developed by linking baseline risk factors in population survey data to a validated population-based diabetes registry to ascertain diabetes diagnosis during follow-up. Specifically, the model was developed in a cohort of 19 861 individuals without diabetes from the province of Ontario who were followed between 1996 and 2005, and was validated in two external cohorts in the provinces of Ontario (n=26 465) and Manitoba (n=9899), as well as across ethnic groups[7 9 14] and in a representative cohort of First Nations people living in Ontario First Nations communities.[15] The algorithm coefficients were updated with more recent data from an Ontario cohort (n=69 606), with follow-up until 2011.[7] The updated DPoRT model has demonstrated high overall predictive performance, good discrimination (C=0.77) and calibration (H-L X2 ≤20).[7] Full details of development and validation can be found from a previous study.[7 9] The DPoRT model is shown in online supplemental table S1. In the DPoRT model, diabetes risk is strongly related to body mass index (BMI) and age. Ethnicity, hypertension, and education are also important risk factors in the model for men and women. For men, smoking, heart disease and income are important independent risk factors; for women, immigrant status is an important predictor for diabetes risk.

In this study, we used population survey data from the USA to externally validate DPoRT for this population. In designing and reporting this study, we adhered to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis reporting guidelines.[16]

Data source

The study cohorts were created using the National Health Interview Survey (NHIS). The continuous NHIS of the National Center for Health Statistics (NCHS) is a complex, multistage probability sample of the US non-institutionalized civilian population that has been conducted every year since 1957.[17] The NHIS collects data through personal household interviews conducted by interviewers employed and trained by the US Bureau of the Census according to procedures specified by NCHS.

The survey provides information on the health of the US population, including information on the prevalence and incidence of disease, the extent of disability, and the use of healthcare services. NHIS was chosen over other nationally represented surveys conducted in the USA, notably the National Health and Nutrition Examination Survey, because it is conducted continuously on an annual basis and is the largest representative health survey to assess diabetes incidence, providing estimates for the incidence of type 2 diabetes in any given year and at the state-level.

Publicly available survey weights were generated for all continuous NHIS survey cycles by the Centers for Disease Control and Prevention (CDC) enabling the generation of estimates representative of the US adult population. Survey weights are necessary to account for non-response and oversampling.

Study population

The external validation cohort consisted of adult respondents to the 2009 NHIS, aged 20–84 years, not pregnant, and without self-reported diabetes at the time of interview (n=23 477). A separate cohort consisting of pooled annual cycles of the 2009–2018 NHIS was used to calculate observed incident diabetes cases to which the 10-year DPoRT diabetes incidence predictions could be compared with. This cohort consisted of respondents aged 20–84 years, who were not pregnant at the time of interview (n=293 327). The 2018 NHIS cycle was used as the DPoRT application cohort to which we estimated 10-year diabetes risk projections and modeled scenarios, excluding respondents aged 20–84 years, not pregnant,

and without self-reported diabetes at the time of interview (n=21 187).

Diabetes outcome

We estimated the number of observed incident cases of diabetes by pooling together the individual estimates from each annual NHIS cycle from 2009 to 2018. In each survey, all sampled adults were asked to answer: "*Have you ever been told by a doctor or health professional that you have diabetes?*" To exclude gestational diabetes, women were asked whether they had been told they had diabetes other than during pregnancy, and those reporting 'yes' were excluded from the sample. Adults who reported being diagnosed with diabetes were then asked at what age they were diagnosed. We followed the methodology used by the US CDC to identify incident cases.[18] Incident cases were identified by subtracting the age at which the respondent was diagnosed from their current age at the time of the survey. Adults who had a value of zero were identified as incident cases. Furthermore, to account for having a birthday during the first year of the diabetes diagnosis, half of the adults who had a value of 1 were classified as incident cases, using random selection from a uniform distribution on the interval (0, 1). Random selection with uniform distribution ensures that each individual has an equal chance of being selected. The annual estimates were weighted to estimate the US adult population using the provided NHIS survey weights in the public use datasets.

Self-reported diabetes in the NHIS has been found to have a high concordance with claims-based identified diabetes[19]; however, not all people who are diagnosed with diabetes will self-report that they have the disease. This phenomenon may be due to a variety of reasons, including not being properly informed about the diagnosis from a health professional, not understanding the term when presented to them, disagreeing with the diagnosis itself, believing they are 'cured' because they are managing the disease appropriately, or hiding the diagnosis because of the stigma that persists about diabetes.[20] In a validation study comparing self-reported diabetes from the Canadian Community Health Survey with a linked registry of physician-diagnosed diabetes, it was found that one in four people with physician-diagnosed diabetes did not self-report having the disease.[20] To account for this underestimation, we applied a 25% correction factor to the observed estimate of incident diabetes cases. This became our reference standard of observed incidence from 2009 to 2018 against which the predictive models were compared.

Predictors

DPoRT uses the following predictors: sex, age, BMI, ethnicity (white ethnicity, other ethnicities), education (less than or secondary graduation, some or post-secondary graduation), immigrant status (immigrant, non-immigrant), prior diagnosis of hypertension, prior diagnosis of heart disease, household income quintile, and smoking (current smoker, non-smoker). BMI is categorized as <23.0 kg/m^2, 23.0–24.9 kg/m^2, 25.0–29.9 kg/m^2, 30.0–34.9 kg/m^2, and ≥35.0 kg/m^2. BMI is included in the model as an interaction with sex-specific age groupings (<45, 45–64, ≥65 years for females) and (<45, ≥45 years for males). Missing values for family income were imputed using multiple imputation; the imputed values were provided in the NHIS.[21] All predictors were defined the same as in the original development cohort.[9]

External validation of DPoRT

External validation of DPoRT was performed by applying the model to the validation cohort to estimate the expected number of diabetes incident cases over the next 10 years. This model, referred to as the original DPoRT, used the same intercept and coefficients as developed and validated in the Canadian population (online supplemental table S1). Individuals with missing values for predictors were excluded (1.6% of sample, n=376), except for missing BMI, which is maintained in the DPoRT model. Predictive performance was assessed by comparing the estimate for the predicted number of incident cases from the original DPoRT with our corrected estimate of observed number of incident diabetes cases, in addition to measures of discrimination and calibration.

We assessed whether model updating could improve DPoRT's predictive performance. We implemented three updating methods: intercept recalibration, logistic recalibration, and model extension. Intercept recalibration involves updating the model intercept, which can improve calibration-in-the-large, which is the average predicted risk compared with the observed outcomes.[22] Re-estimating the model intercept ensures that predictions of the model are on average correct. We re-estimated the model intercept by fitting a logistic regression model with the linear predictor of the original DPoRT model (log of DPoRT risk) as an offset variable. An offset variable is a predictor with a regression coefficient fixed at unity. The second approach, logistic recalibration, aims to correct miscalibration of the original model's linear predictor. For this approach, the model's intercept is recalibrated and all coefficients are adjusted using a common factor. To perform logistic recalibration, we re-estimated the model by fitting a logistic regression model with the linear predictor of the original DPoRT model (log of DPoRT risk) as the only variable.[22]

The third approach, model extension, involves recalibrating the intercept and overall calibration slope, while including additional predictors in the model.[22] For model extension, we performed logistic recalibration and included additional predictors by fitting a logistic regression model with the linear predictor of the original DPoRT model (log of DPoRT risk) and additional variables that could potentially add additional predictive

Table 1 Weighted distribution of baseline characteristics by sex in the external validation cohort*

Risk factor	External validation cohort	
	Male N=10 525	Female N=12 952
Age (years)	51.7	47.4
Age group (years)		
<45	52.3	48.3
45–64	35.5	36.3
≥65	12.2	15.3
Ethnicity†		
White ethnicity	68.7	69.0
Other ethnicities	31.1	30.8
Missing	0.2	0.2
Education		
Less than postsecondary	41.0	38.0
Some or postsecondary graduation	58.4	61.4
Missing	0.6	0.5
Body mass index (kg/m^2)		
<23	13.7	27.7
23.0–24.9	15.5	15.5
25.0–29.9	43.2	28.2
30.0–34.9	17.9	14.1
≥35.0	8.1	10.0
Missing	1.7	4.5
Smoking status		
Current smoker	24.2	18.9
Non-smoker	75.4	80.8
Missing	0.4	0.2
Hypertension		
Yes	25.5	25.2
No	74.4	74.8
Missing	0.09	0.04
Heart disease		
Yes	10.8	9.2
No	89.1	90.7
Missing	0.1	0.1

*Numbers are weighted percentages using weights by the Centers for Disease Control and Prevention.
†Ethnicity is self-described by survey respondents. Others include Hispanic, black, Asian, American Indian/Alaskan Native, other race category, or those who self-identified with multiple ethnicities.

information. We considered population-specific drivers of type 2 diabetes risk that are potentially more salient in the US population, specifically insurance coverage (yes, no) and the following operationalization of ethnicity: white (non-Hispanic), black (non-Hispanic), Hispanic, Asian, American Indian/Alaskan Native, and other. Previous studies have documented elevated diabetes risk among certain ethnic and racial groups in the USA,[23 24] as well as the importance of access to care and insurance coverage on preventing adverse health outcomes, including diabetes.[25]

The probabilities for each updated model were computed using the formula: probability=exp(logit)/(1+(exp(logit))), in which the logit is the sum of the regression coefficients multiplied by their respective predictor variable values.

Development versus validation cohort

There are slight differences in eligibility criteria and outcome definition between the original development cohort and the current validation cohort. Specifically, the original DPoRT model was developed and validated for the adult population 20 years and older. In the US validation, we excluded individuals 85 years of age or older due to an inability to identify incident diabetes cases in this age group (ie, the NHIS categorizes age of respondents 85 years or older into a top code of 85 rather than providing single years of age). As well, the outcome definition used for the original DPoRT model was physician-diagnosed diabetes determined from linked validated registries created from administrative data sources.[26]

Statistical analysis

The predictive performance of the original DPoRT and updated versions of the model were assessed by discrimination and calibration. Discrimination is the ability of the model to differentiate between those who will and will not develop diabetes.[27] Discrimination was measured using a C-statistic, a rank order statistic for predictions against true outcomes, analogous to the area under the receiver operating characteristic curve—a value of 1.0 representing perfect discrimination and 0.5 representing no discrimination.[28] Calibration describes how well the predicted probability of disease agrees with the observed outcomes.[27] We assessed calibration graphically using calibration plots and visually comparing the observed and predicted probability of diabetes across the spectrum of predicted risk. A perfectly calibrated model has a calibration line where the points fit a 45-degree diagonal line. Overall performance of the model was measured by the Brier score which calculates the average prediction error. The Brier score measures the accuracy of predictions by calculating the squared difference between the outcome and predictions.[27] It is a measure of overall agreement between observed and predictive risk with values between 0 and 1, where a score of 0 indicates a perfect model.[29]

We compared the predictive performance of the updated models against the original model, and determined that the original DPoRT model was more accurate than the updated versions when considering overall performance, discrimination and calibration (see

'Results' section). In particular, model extension with ethnicity and insurance coverage worsened discrimination, suggesting that these variables did not add predictive information to improve DPoRT's performance. Given that the original DPoRT model was determined to have the best fit, we applied the original DPoRT model to the NHIS 2018 cycle (DPoRT Application Cohort) to estimate each individual's 10-year risk of developing diabetes. Diabetes risk estimates were averaged across all respondents of the study population to determine overall population-level diabetes risk and risk across subgroups of the population. The number of new diabetes cases was estimated by multiplying the average risk by the population size.

Finally, to demonstrate applicability, we used the best performing model (DPoRT original) to model the population-benefit of a community-wide and high-risk prevention approach on diabetes incidence from 2018 to 2028. The *community-wide prevention scenario* was defined as an intervention for the general population that improves walkability of built environments, which results in small changes in a large portion of the population. We assumed this intervention would achieve a 4% decrease in the prevalence of overweight and 8% decrease in the prevalence of obese, based on a study that estimated differences in the prevalence of overweight and obesity in medium-high versus low walkability areas.[30] We modeled this effect by randomly selecting 4% of individuals in the overweight BMI category (BMI 25.0–29.9) and 8% of individuals in the obese BMI category (BMI ≥30.0) and assigning them the DPoRT risk factor coefficient corresponding to the normal BMI weight category (BMI 23.0–24.9). The *high-risk prevention scenario* was defined as rigorous diet and physical activity promotion programs targeted to adults at increased risk for type 2 diabetes. We assumed this intervention would achieve a 40% relative risk reduction in type 2 diabetes incidence, based on a systematic review of single-group and comparative studies that assessed the effectiveness of diet and physical activity prevention programs on diabetes onset.[31] We defined the target group as the top 10% of the population with highest diabetes risk, as estimated by DPoRT. We assumed that only a small portion of the target group would participate in this type of rigorous intervention, therefore, we applied coverage restrictions by randomly selecting 20% of the target group. Finally, we modeled a *combined approach* in which the community-wide and high-risk prevention scenarios described above were implemented simultaneously.

For each scenario, we applied the risk reduction to the target group and compared the estimated 10-year diabetes risk and cases corresponding to each scenario against the baseline. Population benefit was defined as the absolute number of cases prevented and the absolute risk reduction corresponding to each prevention scenario. All analyses were weighted using sampling weights provided by the CDC and were conducted using the appropriate survey procedures where necessary using SAS V.9.4.

RESULTS

Cohort characteristics

Weighted baseline characteristics for the validation cohort are presented in table 1. The external validation cohort had a similar age and smoking distribution, compared with individuals in the original development cohort.[9] However, there was a larger proportion of individuals with other ethnicities, less than postsecondary education, BMI ≥35, hypertension and heart disease, compared with the original development cohort.[9] Ten-year diabetes incidence was similar between the original development cohort (10.0%) and NHIS validation cohort (10.2%).

Model performance

The observed cumulative number of incident cases from 2009 to 2018 was 15 414 818. The corrected estimate accounting for the discrepancy between self-reported and physician-diagnosed type 2 diabetes was 19 268 523 new cases over 10 years. Online supplemental table S2 compares the observed estimate of incident cases with the predicted cases from the original DPoRT and updated models. The originally validated DPoRT more accurately predicted the 10-year number of incident cases for the total population and across sex and ethnicity, compared with the updated models. Specifically, DPoRT original accurately predicted overall cases within 1%.

DPoRT original demonstrated good discriminative ability (C-statistic=0.778 (males); 0.787 (females)). Discrimination was similar in DPoRT models updated with intercept recalibration and logistic recalibration, but was poor in updated models using model extension (figure 1). DPoRT original Brier scores were 0.070 (males) and 0.073 (females), indicating adequate overall model performance. Brier scores did not improve with DPoRT model updating (figure 1).

Figure 1 shows the calibration plots for the observed and predicted probabilities. Both the male and female original DPoRT models were well calibrated across the spectrum of risk, with the exception of an underestimation at the extremes of risk. Model updating with intercept recalibration and logistic recalibration (male model) universally classified the population as having a low-to-moderate probability of diabetes. Logistic recalibration for the female model classified the population as having moderate-to-high risk. Updated DPoRT models with extension showed poor calibration.

DPoRT application

The original DPoRT predicted 10-year diabetes risk and number of new cases from 2018 to 2028 in the US population is shown in figure 2. DPoRT predicted a 10-year risk of 10.2%, corresponding to 21 076 000 new cases. Diabetes risk estimates increase with age and BMI in a linear fashion, and the number of predicted diabetes cases is highest in those aged 45–60 years and among those with a BMI ≥30 kg/m^2. Although non-Hispanic whites have the lowest 10-year risk by ethnicity, because they represent the largest segment of the population, this translates

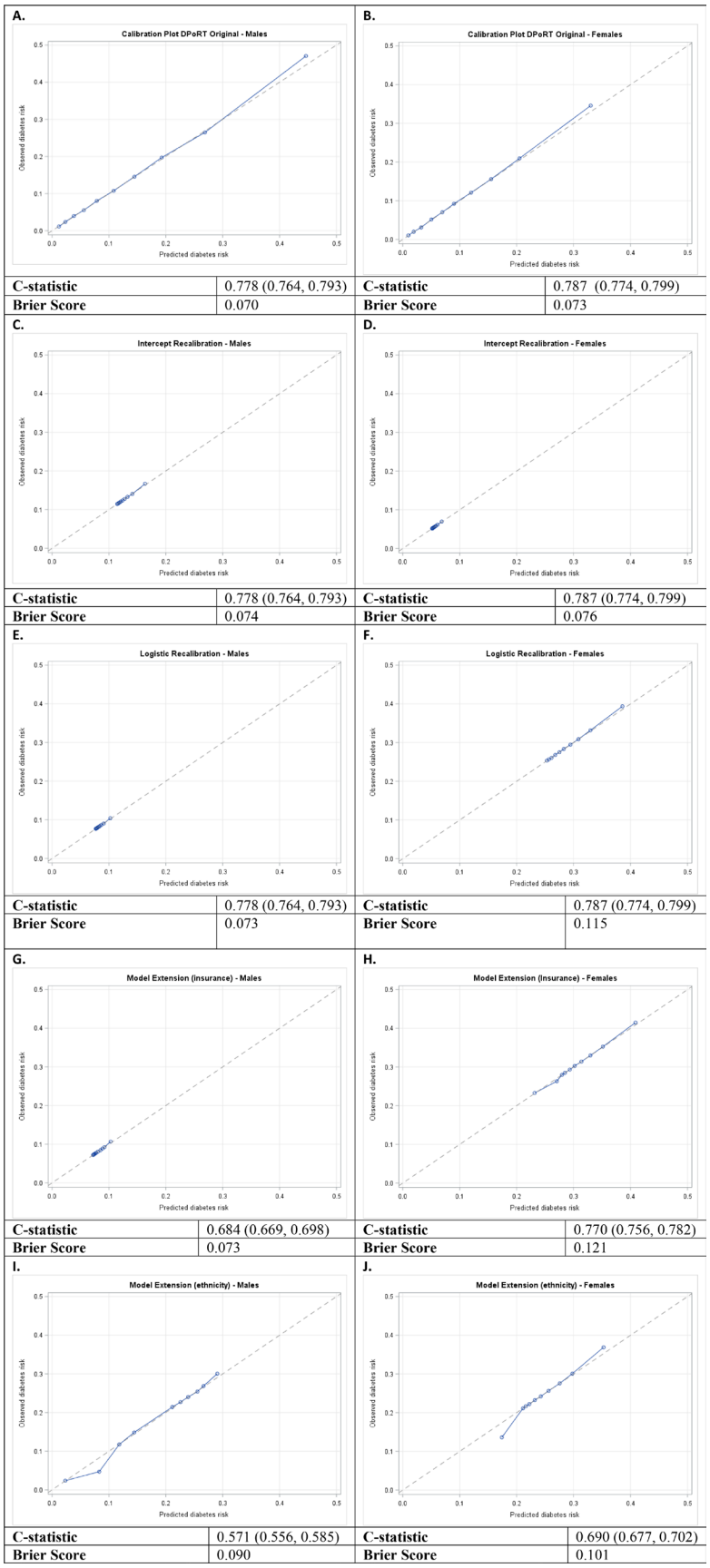


Figure 1 Diabetes Population Risk Tool (DPoRT) predictive performance for the original and updated models in the US population: (A) DPoRT original (male); (B) DPoRT original (female); (C) intercept recalibration (male); (D) intercept recalibration (female); (E) logistic recalibration (male); (F) logistic recalibration (female); (G) model extension (insurance) (male); (H) model extension (insurance) (female); (I) model extension (ethnicity) (male); (J) model extension (ethnicity) (female).

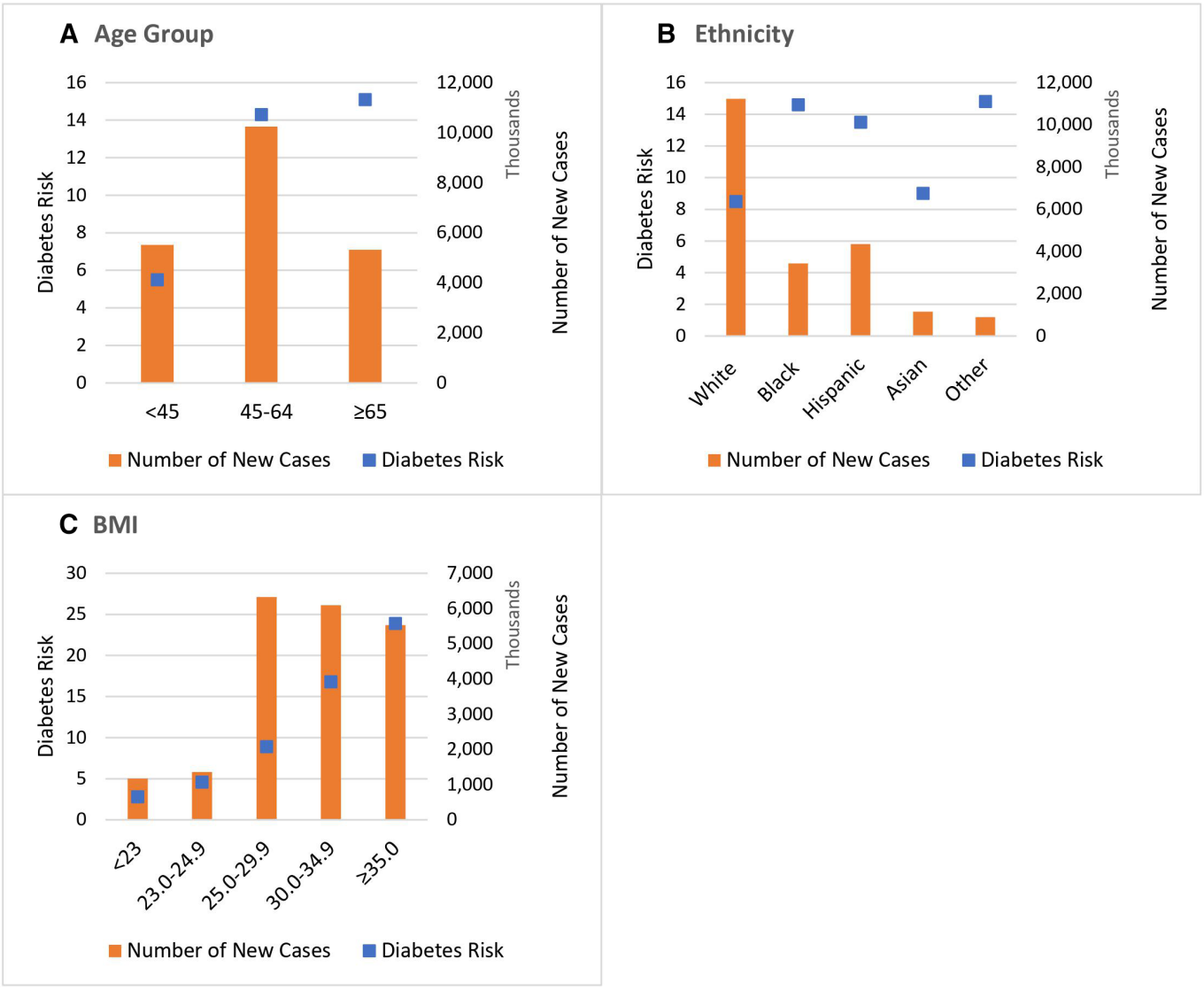


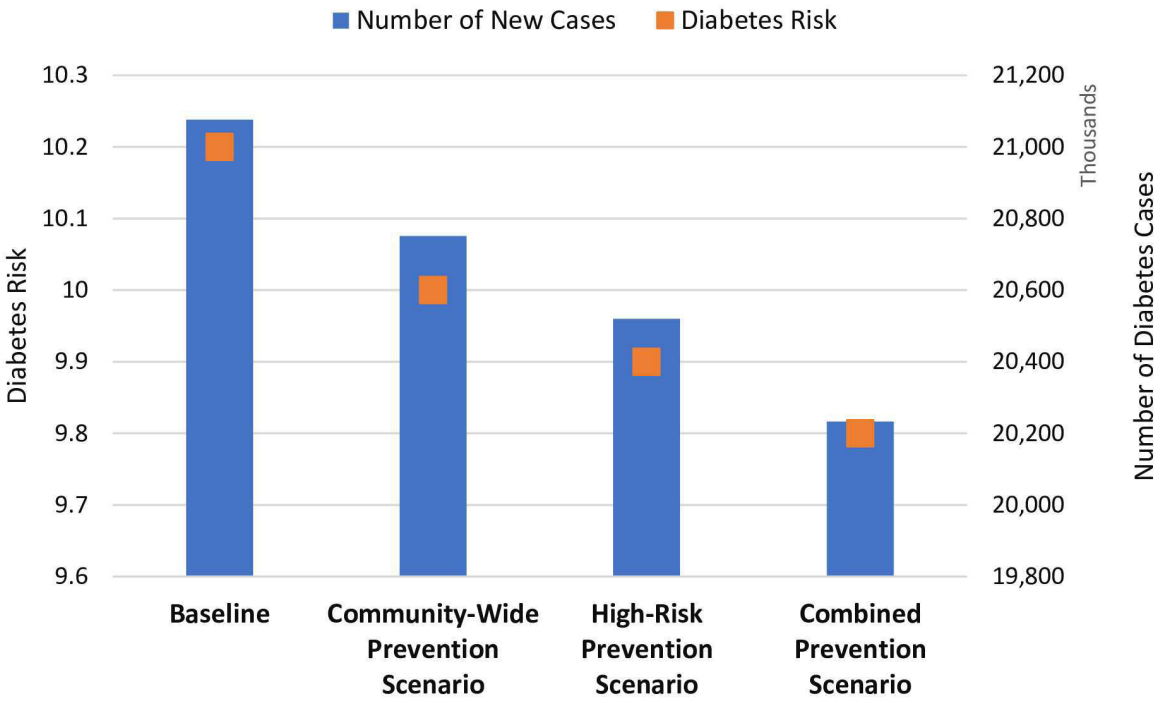
Figure 2 10-Year Diabetes Population Risk Tool estimated diabetes risk and new cases (2018–2028) in the US population according to: (A) age group; (B) ethnicity; (C) body mass index (BMI).

into a high number of expected new cases. The estimated population benefit of the three prevention strategies we modeled is shown in figure 3. The community-wide prevention scenario estimated an absolute risk reduction of 0.2%, corresponding to 325 000 cases prevented. The targeted high-risk prevention scenario estimated an absolute risk reduction of 0.3% and 557 000 cases prevented over 10 years. The combination of community-wide and high-risk prevention strategy was estimated to result in a 0.4% absolute risk reduction in the population and 843 000 cases prevented.

DISCUSSION

This study externally validated DPoRT for the US population. Our model assessment results indicated that the model originally developed in the Canadian population and validated in multiple provinces has adequate predictive performance in the US population, when considering discrimination and calibration. Adjusting the DPoRT models was unable to accommodate a large spectrum of probabilities. Predictive performance also did not improve with model updating using extension demonstrating that the originally validated model performs well in the US population.

Our study demonstrates that diabetes risk can be accurately predicted at the population level using self-reported measures readily available in population health surveys. DPoRT was previously validated in the Canadian context using linkages between survey data with a validated administrative data algorithm for diabetes incidence.[26] In many jurisdictions, linkages with survey and health administrative data either do not exist or are not readily accessible, reflecting a barrier to externally validating existing population risk models.[32] For example, in the USA, the linked NHIS and administrative Medicare data are accessible only under restricted use for approved research projects. Given that survey linkages are not always a



Absolute Risk reduction from baseline (%)	0.2	0.3	0.4
Number of diabetes cases prevented (thousands)	325	557	843

Figure 3 Predicted diabetes risk and cases in the US population at baseline and by prevention scenario (2018–2028).

viable option, this study establishes a robust method for using publicly available population survey data to assess the generalizability of existing population risk models, such as DPoRT, for new settings. Given the importance of ethnicity on developing diabetes, in applying DPoRT in new populations that are multiethnic, it is useful to consider whether a difference in the population's ethnic composition impacts predictive performance. DPoRT includes a non-specific category of ethnicity (ie, white and other ethnicities), with all other ethnicities having a higher risk for developing diabetes. We found that in the current US validation cohort, updating DPoRT with a predictor that included detailed ethnic information did not improve predictive performance, as was also found in a similar study in the Canadian context.[14]

Baseline risk assessment is important for population health decision-making.[8] This study demonstrated the utility of DPoRT for population health assessment and modeling of diabetes prevention scenarios. Our modeling results found that the greatest population benefit, relative to diabetes cases averted, are projected from combining community-wide diabetes prevention approaches with targeted interventions for high-risk groups. Our findings align with a previous study that used a dynamic modeling approach and estimated that combined prevention strategies would be most effective at reducing diabetes incidence,[33] and adds to this research by accounting for age, sex and ethnic group, which are predictors in the DPoRT model. Our findings support current strategies undertaken for high-risk populations through the National Diabetes Prevention Program, and underscore a need for wide-scale implementation of whole-population approaches to meaningfully reduce diabetes burden.[34]

The results should be interpreted in the context of study limitations. The self-reported measures used as DPoRT predictor variables may be subject to reporting error. However, DPoRT was designed to be applied to self-reported data, and unless survey structure or reporting patterns are different across populations, therefore unlikely to meaningfully affect model performance. Specifically, the NHIS uses reliable methods and well-established exposure questions, thus reporting errors are unlikely to differ significantly from the Canadian data source to which DPoRT was previously validated (ie, Canadian Community Health Survey). In addition, DPoRT is validated to predict physician-diagnosed diabetes; the estimates exclude people with undiagnosed diabetes.

Undiagnosed diabetes is estimated to account for 2.8% of the US population, contributing to 21.4% of total type 2 diabetes prevalence (diagnosed plus undiagnosed).[18] Our modeling scenarios were applied with assumptions regarding uptake and effectiveness. To model the community-wide prevention scenario, we applied an evidence-based risk reduction based on an Australian study, which assumes applicability to the US context. In modeling the high-risk prevention scenario, we applied intervention coverage restrictions, but assumed uptake and adherence is uniform across population groups. Thus, our estimates on population-benefit may be overestimated if uptake and adherence is lower in population subgroups, especially those with high risk.

Conclusion

This study externally validated DPoRT using nationally representative data from the USA, and established a methodology that can be used with population survey data to assess the predictive performance of population risk tools in new settings. We demonstrated DPoRT's applicability for population health assessment and modeling population benefit of diabetes prevention strategies. Our projections indicate a need to invest in both community-wide and targeted primary prevention efforts to curb the diabetes burden in the USA.

Acknowledgements This research uses data from the National Health Interview Survey (NHIS), which is conducted annually by the National Center for Health Statistics, Centers for Disease Control and Prevention. The data were accessed through the NHIS Public Use File (2009–2018).

Contributors Conceptualization and methodology: KK, CT, LCR; formal analysis: KK, CT, EN; writing—original draft preparation: KK; writing—review and editing: CT, EN, LCR; funding acquisition: LCR. All authors have read and agreed to the published version of the manuscript. LCR is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Funding This research was supported by the Canada Research Chair held by LCR in Population Health Analytics (CRC 950-230702) and by the Banting and Best Diabetes Centre, University of Toronto (EN).

Disclaimer The findings and conclusions presented in this research are solely those of the authors and do not necessarily represent the views of the National Center for Health Statistics or the Centers for Disease Control and Prevention.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study was approved by the University of Toronto (protocol #44476). The NHIS protocol was approved by the National Center for Health Statistics (NCHS) institutional review board, and verbal consent for survey participation was obtained from all participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. The NHIS data are available from www.cdc.gov/nchs/nhis/index.htm.

ORCID iD

Laura C Rosella http://orcid.org/0000-0003-4867-869X

REFERENCES

1 Cho NH, Shaw JE, Karuranga S, *et al*. IDF diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract* 2018;138:271–81.
2 Zhou B. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. *Lancet* 2016;387:1513–30.
3 Wang L, Li X, Wang Z, *et al*. Trends in prevalence of diabetes and control of risk factors in diabetes among US adults, 1999-2018. *JAMA* 2021;326:1–13.
4 Liu J, Ren Z-H, Qiang H, *et al*. Trends in the incidence of diabetes mellitus: results from the global burden of disease study 2017 and implications for diabetes mellitus prevention. *BMC Public Health* 2020;20:.:1415.
5 Konchak JN, Moran MR, O'Brien MJ, *et al*. The state of diabetes prevention policy in the USA following the affordable care act. *Curr Diab Rep* 2016;16:55:1–12.:.
6 Manuel DG, Rosella LC, Tuna M, *et al*. Effectiveness of community-wide and individual high-risk strategies to prevent diabetes: a modelling study. *PLoS One* 2013;8:e52963.
7 Rosella LC, Lebenbaum M, Li Y, *et al*. Risk distribution and its influence on the population targets for diabetes prevention. *Prev Med* 2014;58:17–21.
8 Manuel DG, Rosella LC, Hennessy D, *et al*. Predictive risk Algorithms in a population setting: an overview. *J Epidemiol Community Health* 2012;66:859–65.
9 Rosella LC, Manuel DG, Burchill C, *et al*. A population-based risk algorithm for the development of diabetes: development and validation of the diabetes population risk tool (Dport). *J Epidemiol Community Health* 2011;65:613–20.
10 Rosella LC, Bornbaum C, Kornas K, *et al*. Evaluating the process and outcomes of a knowledge translation approach to supporting use of the diabetes population risk tool (Dport) in public health practice. *Can J Program Eval* 2018;33:21–48.
11 Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. *J Clin Epidemiol* 2016;69:245–7.
12 Collins GS, de Groot JA, Dutton S, *et al*. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;14:40:1–11.:.
13 Binuya MAE, Engelhardt EG, Schats W, *et al*. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. *BMC Med Res Methodol* 2022;22:316.
14 Rosella LC, Mustard CA, Stukel TA, *et al*. The role of Ethnicity in predicting diabetes risk at the population level. *Ethnicity & Health* 2012;17:419–37.
15 Rosella LC, Kornas K, Green ME, *et al*. Characterizing risk of type 2 diabetes in first nations people living in first nations communities in Ontario: a population-based analysis using cross-sectional survey data. *Cmajo* 2020;8:E178–83.
16 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. *Circulation* 2015;131:211–9.
17 Parsons VL. US Department of Health and Human Services, Centers for Disease Control and Prevention. *Design and estimation for the national health interview survey, 2006-2015*. National Center for Health Statistics, 2014.
18 Centers for Disease Control and Prevention. National diabetes Statistics report 2020: estimates of diabetes and its burden in the United States; 2020.
19 Day HR, Parker JD, US Department of Health and Human Services, Centers for Disease Control and Prevention. Self-report of diabetes and claims-based identification of diabetes among Medicare beneficiaries; 2013
20 Shah BR, Manuel DG. Self-reported diabetes is associated with self-management behaviour: a cohort study. *BMC Health Serv Res* 2008;8:142.

21 Schenker N, Raghunathan TE, Chiu P-L, *et al*. Multiple imputation of missing income data in the national health interview survey. *J Am Stat Association* 2006;101:924–33.
22 Steyerberg EW, Borsboom GJJM, van Houwelingen HC, *et al*. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23:2567–86.
23 Dias J, Echeverria S, Mayer V, *et al*. Diabetes risk and control in multi-ethnic US immigrant populations. *Curr Diab Rep* 2020;20:73:1–11.:.
24 Hassan S, Gujral UP, Quarells RC, *et al*. Disparities in diabetes prevalence and management by race and Ethnicity in the USA: defining a path forward. *Lancet Diabetes Endocrinol* 2023;11:509–24.
25 Ekperi LI, *et al*. Health insurance coverage among diabetic adults from three major ethnic groups in the United States. *Ethnicity Disease* 2012;22:486–91.
26 Hux JE, Ivis F, Flintoft V, *et al*. Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. *Diabetes Care* 2002;25:512–6.
27 Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* 2010;21:128–38.
28 Harrell FE. Regression modeling strategies. In: *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis* 608. New York, NY: Springer, 2001:
29 Rufibach K. Use of brier score to assess binary predictions. *J Clin Epidemiol* 2010;63:938–9;
30 Mayne DJ, Morgan GG, Jalaludin BB, *et al*. Area-level Walkability and the geographic distribution of high body mass in Sydney, Australia: a spatial analysis using the 45 and up study. *Int J Environ Res Public Health* 2019;16:664.
31 Balk EM, Earley A, Raman G, *et al*. Combined diet and physical activity promotion programs to prevent type 2 diabetes among persons at increased risk: a systematic review for the community preventive services task force. *Ann Intern Med* 2015;163:437–51.
32 Siontis GCM, Tzoulaki I, Castaldi PJ, *et al*. External validation of new risk prediction models is infrequent and reveals worse Prognostic discrimination. *Journal of Clinical Epidemiology* 2015;68:25–34.
33 Gregg EW, Boyle JP, Thompson TJ, *et al*. Modeling the impact of prevention policies on future diabetes prevalence in the United States: 2010–2030. *Popul Health Metr* 2013;11:18:1–9.:.
34 Gruss SM, Nhim K, Gregg E, *et al*. Public health approaches to type 2 diabetes prevention: the US national diabetes prevention program and beyond. *Curr Diab Rep* 2019;19:78:78.:.