The role of ethnicity in predicting diabetes risk at the population level

Laura C. Rosella,a,b,c*, Cameron A. Mustardc,e, Therese A. Stukelb,d, Paul Coreye, Jan Huxb,d, Les Roosf and Douglas G. Manuelb,g,h

aPublic Health Ontario, Toronto, Ontario, Canada; bInstitute for Clinical Evaluative Sciences, Toronto, Ontario, Canada; cDalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; dDepartment of Health Policy, Management, and Evaluation, University of Toronto, Toronto, Ontario, Canada; eInstitute of Work and Health, Toronto, Ontario, Canada; fUniversity of Manitoba, Winnipeg, Manitoba, Canada; gOttawa Hospital Research Institute, Ottawa, Ontario, Canada; hStatistics Canada, Ottawa, Ontario, Canada

(Received 20 September 2010; final version received 28 November 2011)

Background. The current form of the Diabetes Population Risk Tool (DPoRT) includes a non-specific category of ethnicity in concordance with publicly data available. Given the importance of ethnicity in influencing diabetes risk and its significance in a multi-ethnic population, it is prudent to determine its influence on a population-based risk prediction tool.

Objective. To apply and compare the DPoRT with a modified version that includes detailed ethnic information in Canada’s largest and most ethnically diverse province.

Methods. Two additional diabetes prediction models were created: a model that contained predictors specific to the following ethnic groups – White, Black, Asian, south Asian, and First Nation; and a reference model which did not include a term for ethnicity. In addition to discrimination and calibration, 10-year diabetes incidence was compared. The algorithms were developed in Ontario using the 1996–1997 National Population Health Survey (N = 19,861) and validated in the 2000/2001 Canadian Community Health Survey (N = 26,465).

Results. All non-white ethnicities were associated with higher risk for developing diabetes with south Asians having the highest risk. Discrimination was similar (0.75–0.77) and sufficient calibration was maintained for all models except the detailed ethnicity models for males. DPoRT produced the lowest overall ratio between observed and predicted diabetes risk. DPoRT identified more high risk cases than the other algorithms in males, whereas in females both DPoRT and the full ethnicity model identified more high risk cases. Overall DPoRT and full ethnicity algorithms were very similar in terms of predictive accuracy and population risk.

Conclusion. Although from the individual risk perspective, incorporating information on ethnicity is important, when predicting new cases of diabetes at the population level and accounting for other risk factors, detailed ethnic information did not improve the discrimination and accuracy of the model or identify significantly more diabetes cases in the population.

Keywords: diabetes; risk prediction; ethnicity
Introduction

Planning for health care and public health resources needed to address the significant burden of diabetes (Wild et al. 2004) is an important aspect of population health management, which can be informed by robust prediction tools, such as the Diabetes Population Risk Tool (DPoRT; Rosella et al. 2010). This tool can aid policy-makers, planners, and physicians by providing reliable estimates of the upcoming diabetes epidemic. In addition, the effectiveness of widespread prevention strategies can be improved by knowing which groups to target and how extensive a strategy is needed to stabilize or reduce the number of new cases.

Risk prediction tools for estimating disease risk are common in clinical settings and are used for clinical decision-making (Anderson et al. 1991). One of the limitations of clinical risk prediction tools for population prediction is the reliance on physical measurements or special risk questions, such as fasting blood sugar (Ito et al. 1996, Eddy and Schlessinger 2003, Hanley et al. 2003;) or diabetes family history (Herman et al. 1995, Lindstrom and Tuomilehto 2007) in the case of diabetes. At the population level, these measurements are often not easily, accurately, or systematically captured. One of the key attributes of DPoRT is its accessibility to a broad audience. This is achieved by using data from surveys that are publicly available. In these surveys, detailed ethnic information, though often collected, is not publicly reported. In Canada, ethnic information from the surveys are reported publicly as 'white/non-white,' and thus this form for ethnicity was used in DPoRT in order to ensure that the tool can be applied to publicly available data in Canada.

There is growing evidence that certain ethnic groups are at increased risk for developing type 2 diabetes. Globally, non-European populations have a higher proportional burden of type 2 diabetes compared to the other regions of the world (World Health Organization 1998). The highest diabetes rates in the world are seen in aboriginal population, including those in Australia(Odea 1991, Odea et al. 1993), USA(Mokdad et al. 2001, Pavkov et al. 2007), and Canada(Harris et al. 1997; Young et al. 2000). Throughout the world, those of south Asian decent are shown to carry an increase burden of type 2 diabetes compared with both non-white and white ethnicities (Ramachandran et al. 1997a, b, Abate and Chandalia 2001). Data from Ontario demonstrate that overall immigrant and ethnic minority populations suffer from a higher burden of diabetes and its complications (Manuel and Schultz 2003). The importance of ethnicity when considering those at high risk for developing diabetes in the clinical setting has been emphasized through diabetes guidelines that recommend people of Aboriginal, Hispanic, south Asian, Asian, or African descent should be targeted for screening (Calonge et al. 2008, Norris et al. 2008). Canada’s immigrant population is largely made up of non-white ethnicities. Immigrants account for 18–20% of Canada’s population (Newbold and Danforth 2003), and this percentage is expected to increase over time. Estimates of immigrant populations are as high as 50% for major urban centers such as Toronto.

Although clinically and epidemiologically important risks associated with ethnicity are apparent, it is not clear how omitting ethnic-specific predictors will affect a population-based prediction tool for diabetes. Given the significance of ethnicity in Canada and its important influence on diabetes risk, it may be possible that failing to apply ethnic-specific predictors will reduce the ability to identify high risk groups. In order to have confidence in applying this tool, it needs to be
determined if the inclusion of detailed ethnic predictors will significantly change the performance of DPoRT.

The purpose of this study was to assess the impact of including detailed ethnic information in a prediction algorithm for diabetes. Specifically this study describes the relative benefits to predictive accuracy and model outputs that are gained with the addition of ethnic specific predictors to the model. In addition to informing the application of DPoRT, this work also provides insight into the independent role of ethnicity on diabetes risk once additional risk factors are considered.

Methods

Creation and validation of DPoRT

Development cohort

The study cohort was derived from 23,403 people from Ontario that responded to the 1996/1997 National Population Health Survey conducted by Statistics Canada. In the NPHS, households were selected though a stratified, multilevel cluster sampling of private residences using provinces and/or local planning regions as the primary sampling unit. The survey, conducted via telephone, had an overall 83% response rate and all responses were self-reported. Persons under the age of 20 (n=2,407) and those who had previously diagnosed diabetes or self-reported diabetes were excluded (n=894). Those who were pregnant at the time of the survey were also excluded (n=241), due to the fact that baseline body mass index (BMI) could not be accurately ascertained, and males with missing BMI (n=66) were excluded leaving a total of 19,795 individuals in the final cohort. This cohort is referred to as: ‘derivation cohort (NPHS)’.

Validation cohorts

Diabetes Population Risk Tool was validated in two external cohorts in the overall population (Rosella et al. 2010); however, this validation was not conducted using alternate forms of the algorithm nor specifically among ethnic populations, which require the use of different data sources. One external validation cohort was used in this study to compare the performance of the three risk algorithms. The validation cohort used in this study was derived from the Ontario portion of the 2000/2001 Canadian Community Health Survey (CCHS, Cycle 1.1, N=37,473), a national telephone survey administered by Statistics Canada, known herein as ‘validation cohort (CCHS-2000/2001)’. The target population of the CCHS consisted of persons aged 12 and over resident in private dwellings in all provinces and territories, excepting those living on Aboriginal reserves, on Canadian Forces Bases, or in some remote places. The CCHS included the same self-reported health questions as the NPHS. Like the NPHS, this survey uses a multistage stratified cluster design and provides cross-sectional data representative of 98% of the Canadian population over the age of 12 years, and attained an 80% overall response rate (Statistics Canada 2002, 2003). After the exclusion criteria were applied there were 26,465 individuals in the validation cohort. Five years of follow-data were available in the validation cohort.
Identifying respondents who develop diabetes

Survey data from development and validation cohorts were linked to provincial administrative health care databases that include all persons covered under the government funded universal health insurance plan. The diabetes status of all respondents in Ontario was established by linking persons to the Ontario Diabetes Database (ODD). The ODD contains all physician diagnosed diabetes patients in Ontario identified since 1991. The database is created using hospital discharge abstracts and physician service claims. A patient is said to have physician diagnosed diabetes if he/she meets at least one of the following two criteria: (1) 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, or (2) a physician services claim with a diabetes diagnosis (code 250) followed within two years by either a physician services claim or a hospital admission with a diabetes diagnosis. Individuals entered the ODD as incident cases when they were defined as having diabetes according to the criteria described earlier. A hospital record with a diagnosis of pregnancy care or delivery close to a diabetic record (i.e., a gestational admission date between 90 days before and 120 days after the diabetic record date), were considered to represent gestational diabetes therefore were excluded. The ODD has been validated against primary care health records as an accurate measure of incidence and prevalence of diabetes in Ontario (sensitivity of 86%, specificity of 97%) (Hux and Ivis 2005, Lipscombe and Hux 2007). Information regarding the vital statistics and eligibility for health care coverage for linked respondents was captured from the Registered Persons Database. The ODD algorithm is applied nationally using provincial administrative registries (known as the National Diabetes Surveillance System) and has been used and validated in several Canadian provinces (Health Canada 2003).

Variable definitions

Variables used in this study were obtained from responses in the NPHS and CCHS, including: age, BMI, presence of chronic conditions diagnosed by a health professional (including hypertension and heart disease), ethnicity, immigration status, smoking status, highest level of achieved education. BMI in kg/m² was used as an indicator of obesity. Derived BMI, calculated by dividing the weight in kilograms by height squared in meters-squared directly from the NPHS, is only calculated for respondents aged 30–64 years; therefore, BMI was calculated using weight and height according to derived variable specification for those who fell outside the age range of 30–64 years (Statistics Canada 1999). Ethnicity was ascertained by the question, ‘To which ethnic or cultural groups do your ancestors belong?’ Individuals were able to classify themselves in the following categories: White, Chinese, south Asian, Black, Filipino, Latin American, Southeast Asian, Arab, west Asian, Japanese, Korean, Aboriginal/First Nation (North American, Indian, Métis or Inuit, or Other). For the purpose of this study, the term ‘ethnicity’ is used to represent race or ethnicity using the previously listed categories given the data available in this survey. Further classification and aggregation of ethnic groups were based on pre-identified groups that are recommended for diabetes screening.
consistent with the diabetes screening guidelines (Berg et al. 2003, Canadian Diabetes Association Clinical Practice Guidelines Expert Committee 2008). White, Black, Asian (Chinese, Japanese, Korean, Filipino), south Asian, and First Nation according to Statistics Canada’s definition (Statistics Canada 2001). Asian combination was based on similarity of diabetes risk. All others were classified in the ‘other’ category, including Southeast Asian, Arab, Latin American, and west Asian due to the very small sample sizes of these populations in the survey. Statistics Canada releases public-use data files of the national health surveys; however, certain variables are suppressed or modified in these files to protect privacy. In the public-use files ethnicity is only categorized as white or non-white, derived from the response to the ethnicity question. The shared population health survey files, which are available at the provincial level, contain more detailed information (including detailed ethnicity). Access to the shared data files is highly restricted, which is why DPoRT was developed using variables from the public-use file rather than the provincial file. In this study the shared file was used in order to allow for both forms of the variable (white/non-white versus five ethnic groups) to be compared.

**Statistical analysis**

**Creation of DPoRT**

The goal in the creation of DPoRT was to create a risk algorithm that would accurately predict diabetes risk with high discrimination and calibration using risk factors that are measured reliably from health survey data. A detailed description of the development of DPoRT can be found elsewhere (Rosella et al. 2010). Briefly, for each cohort member, the probability of physician diagnosed diabetes was assessed from the interview date until censoring for death or end of follow-up using a Weibull accelerated failure time model. Diabetes risk functions were derived separately for men and non-pregnant women above the age of 20 without a prior diabetes diagnosis. Overall risk (predicted probability) of diabetes for each person was calculated by multiplying the individual’s risk factor values by the corresponding regression coefficients, and summing the products (Odell et al. 1994).

All estimates (including coefficient of regressions and variance estimates) incorporated bootstrap replicate survey weights to accurately reflect the demographics of the Canadian population and account for the survey sampling design based on selection probabilities and post stratification adjustments. Variance estimates and 95% confidence intervals were calculated using bootstrap survey weights to account for the survey sample (Yeo et al. 1999, Kovacevic et al. 2008). All statistics were computed using SAS statistical software (version 9.1 SAS Institute Inc, Cary, NC).

**Creation of additional models**

Two additional models were created in the development cohort (NPHS) as described earlier, however, these new models were modified to either include ethnic-specific predictors or remove any ethnic predictors; therefore, in total three prediction models were compared:
(1) DPoRT minus ethnicity – known as ‘no ethnicity’
(2) DPoRT
(3) DPoRT plus detailed ethnic information – known as ‘full ethnicity’

In DPoRT (model (2)) ethnicity is grouped as in the public-use file as white/non-white.

In model (3) ethnicity is broken up into the pre-identified categories consistent with the diabetes screening guidelines White, Black, Asian, south Asian, and First Nation.

Comparison across models

The performance of the model was measured by the discrimination and calibration values in the external validation cohort. Discrimination is the ability to differentiate between those who are high risk and those who are low risk – or in this case those who will and will not develop diabetes given a fixed set of predictor variables. Discrimination was measured using a C-statistic modified for survival data developed by Pencina and D’Agostino (2004), analogous to the area under the ROC curve (Campbell 2004). An ROC curve repeats all possible pairings of subjects in the sample who exhibit the outcome (in this case diabetes) and do not exhibit the outcome and calculates the proportion of correct predictions. This results in an index of resolution of the model. This proportion under the receiving operator curve is equal to the C-statistic which can be used to assess the degree of discrimination – 1.0 being perfect discrimination and 0.5 being no discrimination (Harrell et al. 1996, Harrell 2001, Pencina and D’Agostino 2004). A model with perfect discrimination would perfectly resolve the population into those who get diabetes and those who do not. Calibration or accuracy is a measure of how accurately predicted probabilities closely agree with the observed outcomes. Calibration is unaffected by discrimination, meaning a model can possess good discrimination yet have poor calibration. A model that does not have sufficient calibration will have significant over- or under-estimation of diabetes risk in the overall population and/or within certain subgroups. A model with good accuracy model will maintain reliability across various risk groups and other important subpopulations. Calibration is not an issue if the purpose of the predicted model is only to rank-order subjects (Harrell 2001). Calibration was assessed using a modified version of the Hosmer-Lemenshow $\chi^2$ statistic developed by Nam (Nam 2000, D’Agostino et al. 2001). This statistic is computed by dividing the validation cohort into deciles of predicted risk of diabetes and compared the observed versus predicted risk in each decile using a chi-square statistic. To mark sufficient calibration $\chi^2 = 20$ was used as a cut-off ($p < 0.01$), consistent with D’Agostino’s validation of the Framingham algorithms (D’Agostino et al. 2001). Observed versus predicted cases of diabetes were also compared across ethnicities to examine concordance across ethnic groups using the three algorithms. All discrimination and validation measures were calculated using a 5-year cohort given that 5-years of follow-up data were available in the validation cohort (CCHS-2000/2001).

The policy implications of the three models were assessed by applying the model to the 2000/2001 CCHS baseline data to predict 10-year diabetes incidence rates (10-year prediction) and cases and then to compare these values between the algorithms. The proportion of the population who were identified as high risk for
developing diabetes were also reported and compared across algorithms; where high risk is classified as \( \geq 20\% \) probability of developing diabetes in 10 years. In addition, observed and predicted diabetes cases were calculated across ethnicity to capture performance of the model within ethnic groups.

In order to describe the impact of disagreement between observed and predicted diabetes risk as a function of the proportion of the population where that disagreement exists, an index called the population disagreement index (PDI) was developed. PDI was summarized across ethnic groups and compared between models.

This PDI is defined as follows:

\[
PDI = \sum_{i=0}^{n} \left( \frac{0_i}{P_i} \times P_{pi} \right)
\]

Where \( 0_i/P_i = \) Ratio between observed:predicted in subgroup \( i \), \( P_{pi} = \) Proportion of the population made up of subgroup \( i \), \( i = 1 \ldots n \) where \( n = \) number of subgroups in the population, \( 0 < PDI < \infty \).

The unweighted ratio between observed and predicted were calculated to demonstrate the influence of the distribution of the subgroup (i.e., \( P_{pi} \)). This was calculated by taking the overall ratio summed across ethnic groups, i.e., \( \sum (0_i/P_i) \). A model that predicts fewer cases than observed would have a ratio \( < 1 \) and one that predicts more cases than observed would have a ratio \( > 1 \).

**Results**

The observed 10-year diabetes risk (cumulative incidence rate) in the Ontario development cohort was 8.8\% for males and 7.3\% for non-pregnant females aged 20 years and older at baseline. In addition to non-white ethnicity the other attributes in the model which were previously validated were: BMI, age (and its interactions), hypertension, smoking, heart disease, and immigrant status.

**Ethnicity**

Seventeen percent of the 2000/2001 cohort self-identified as a non-white ethnicity. Adjusted hazard ratios for the ethnic categories in DPoRT and the full ethnicity algorithm are shown in Figure 1. Non-white ethnicity has a hazard ratio of 2.14 95\% CI (1.74, 2.63) in males and 1.71 (1.35, 2.16) in females, adjusted for all other variables in the risk algorithm. In the full ethnicity model, hazard ratios for specific ethnic groups ranged from 1.11 to 3.02 compared to white ethnicity. All non-white ethnic groups were at higher diabetes risk than white ethnicity. South Asian ethnicity had the highest hazard ratio for diabetes in both males and females.
Model performance

All three models showed good discrimination in the development cohort (NPHS) and validation cohort (CCHS 2000/2001) (C-statistic ranging from 0.75 to 0.77). The algorithm with white/non-white ethnicity predictors had slightly higher discrimination versus a model with no ethnicity (i.e., DPoRT versus no ethnicity algorithm) (Table 1). The full ethnicity algorithm achieved the same discrimination as DPoRT for males and females. Sufficient calibration/accuracy ($\chi^2_{H-L} < 20$) was maintained in
the development and validation cohort (CCHS 2000/2001) for all models except the full ethnicity model for males ($\chi^2_{H-L} = 33.9$). All models had a similar ratio of observed to predicted diabetes risk across decile of risk (Figure 2). Of the three risk algorithms DPoRT had the lowest overall average ratio between the observed and predicted (1.09 versus 1.89 without ethnicity and 1.32 with detailed ethnicity in males and 1.03 versus 1.17 and 1.08, respectively, in females). Weighting by population proportion in the PDI calculation significantly reduces the overall disagreement in the population in all three algorithms due to the fact that larger disagreement occurs in smaller proportions of the population (Table 1). Overall the PDI was lower for women than men.

**Policy implications – estimates of 10-year population diabetes risk in Ontario**

In males the overall 10-year predicted diabetes risk ranged from 9.85% for the no ethnicity model, 10.11% in DPoRT and 10.06% in the full ethnicity model. In females average 10-year predicted diabetes risk ranged from 7.83% for the no ethnicity model, 7.95% in DPoRT and 7.97% in the full ethnicity model. There were 9660 more predicted cases in males and 5013 predicted cases in females in DPoRT than with the model without ethnicity. In males, 1409 less cases were predicted in the full ethnicity model compared to DPoRT and in females 934 more cases were predicted in the full ethnicity model compared with DPoRT (Table 1).

Overall, DPoRT appears to identify more cases at high risk for diabetes than the other two algorithms in males, whereas in females both DPoRT and the full ethnicity were substantially different (Table 1). Across decile of risk the number of diabetes cases predicted using DPoRT and full ethnicity algorithms were very similar in both males and females. Observed and predicted diabetes cases across ethnic groups were most similar using DPoRT, particularly among females (Figure 3). The biggest discrepancy was seen among south Asian males in all algorithms. Females did not have the same discrepancy among the south Asian group where even the inclusion of a term for south Asian ethnicity in the full ethnicity algorithm resulted in an underestimate of diabetes cases. Overall, the largest number of cases belongs to the ‘white’ category as they represent the largest ethnic group in this study population.

**Discussion**

The aim of this study was to assess the impact of including detailed ethnic predictors in a population-based risk tool for diabetes. In addition to identifying relative hazards of developing diabetes by ethnicity, this study provides estimates of the predicted number of cases in a provincial population by ethnicity for the next 10-year period. Using a population-based cohort, this study confirmed that those of non-Caucasian descent are at increased risk for developing diabetes and consistent with previous research, hazard ratios were highest among south Asians (Abate and Chandalia 2001). In terms of overall model performance, no additional predictive value was detected when adding detailed ethnic predictors. At the population level, distribution of diabetes risk was similar in the population, particularly between DPoRT and the full ethnicity model. This study suggests that using DPoRT in its current form is sufficient for accurately predicting diabetes cases in ethnically diverse population. The finding that the algorithm to predict diabetes that uses detailed
Table 1. Ten-year risk, predicted new diabetes cases from 2000/2001 to 2010/2011 and ratio of observed and predicted risk of diabetes in the validation cohort (CCHS 2000/2001).

| Model          | Diabetes rate (%) | Median rate (%) | Proportion at high risk | # New cases | Population disagreement Index<sup>a</sup> | Diabetes rate (%) | Median rate (%) | Proportion at high risk | # New cases |
|----------------|-------------------|-----------------|-------------------------|-------------|--------------------------------------------|-------------------|-----------------|-----------------------|-------------|
|                |                   |                 |                         | Unweighted<sup>c</sup> | Weighted<sup>c</sup> |                   |                 |                         | Unweighted<sup>c</sup> | Weighted<sup>c</sup> |
| No ethnicity   | 9.85              | 5.78            | 14.6                    | 386,964     | 1.89                                       | 7.83              | 4.18            | 8.17                  | 321,724     | 1.27                                      | 1.08                                      |
| DPoRT          | 10.11             | 6.50            | 16.0                    | 396,624     | 1.09                                       | 7.95              | 4.74            | 9.16                  | 326,737     | 0.93                                      | 1.07                                      |
| Full ethnicity | 10.06             | 6.74            | 14.5                    | 395,215     | 1.32                                       | 7.97              | 4.69            | 9.32                  | 327,671     | 1.06                                      | 1.09                                      |

Note: DPoRT, Diabetes Population Risk Tool.

<sup>a</sup>The population disagreement index (PDI) is the ratio of the observed and predicted diabetes risk summed across ethnic groups.

<sup>b</sup>High risk for developing diabetes is defined as having a predicted probability of developing diabetes $\geq 20\%$ in 10-years.

<sup>c</sup>The unweighted PDI indicated is the ratio of the observed and predicted diabetes risk summed over ethnic groups without considering the sample size of each group, whereas the weighted PDI weighted the ethnic-specific observed and predicted values by the population size of each ethnic group.
Figure 2. Five-year observed versus predicted number of diabetes cases in the validation cohort (CCHS 2000/2001) by decile of risk for males and females using three algorithms: no ethnicity (No Ethnicity), with ‘white/non-white’ ethnicity (DPoRT), and with detailed ethnic predictors (Full Ethnicity).

aAll male models include terms for age, body mass index (BMI), hypertension, heart disease, smoking status, and education. All female models include terms for age, BMI, hypertension, immigrant status, and education.
Figure 3. Five-year observed versus predicted number of diabetes cases in the validation cohort (CCHS 2000/2001) by ethnicity for males and females using three algorithms*: no ethnicity (No Ethnicity), with ‘white/non-white’ ethnicity (DPoRT), and with detailed ethnic predictors (Full Ethnicity).

*All male models include terms for age, body mass index (BMI), hypertension, heart disease, smoking status, and education. All female models include terms for age, BMI, hypertension, immigrant status, and education.
ethnicity did not significantly differ from one that uses a broad categorization of ethnicity can be explained by two mechanisms involving statistical prediction and population disagreement index (PDI).

The results from this study using a population-based risk prediction tool are similar to other studies with clinical risk functions, which demonstrate good discrimination, even when applied to multiethnic cohorts (D’Agostino et al. 2001, Mann et al. 2010). There are several reasons why a clinically important risk factor may not improve the performance of a prediction tool. Even though a variable is independently associated with an outcome, it may not provide incremental improvements in test characteristics in the context of existing predictors. This phenomenon has also been shown for other clinical predictors and outcomes such as C-reactive protein for cardiac risk prediction (Lloyd-Jones et al. 2006). In fact, research has shown that although a battery of novel risk factors have been developed for the prediction of major coronary heart disease (CHD) events, these novel factors have been generally unimpressive in their ability to improve CHD prediction (Wilson et al. 2005). Furthermore, it has been shown that for a variable to make significant improvements in discrimination (i.e., improvement in AUC from 0.8 to 0.9) its multivariable odds ratio must be 6.9 or greater. Pepe et al. (2004) suggesting that in order for detailed ethnicity to improve the algorithm beyond its current discrimination, the adjusted hazard ratio must be very large in magnitude. Interestingly, the model without ethnicity in any form was not detectably worse in terms of model performance such that discrimination and calibration were only marginally decreased compared to DPoRT. This is likely due to the fact that many of the reasons that ethnicity plays a role in diabetes risk are related to other factors captured in the model. In particular, socioeconomic status, obesity (particularly younger onset of obesity), and other lifestyle factors have been shown to be related to both ethnicity and diabetes risk (Abate and Chandalia 2001). Most importantly, immigration status, captured in the model, may explain a significant amount of the variability in diabetes incidence that is associated with ethnicity. The diminishing return on model performance when adding statistically significant predictors to the model was also noted in the model building process of DPoRT and was one of the reasons that DPoRT maintained good discrimination, even with considerable constraints on variable selection (Rosella et al. 2010).

The population disagreement index (PDI) is an extension of the idea of population attributable risk (PAR). PAR describes the impact of a risk factor on population risk as a function of the prevalence of the exposure and the relative risk of disease (Levin and Bertell 1978). In this study, the estimate of relative risk in PAR is translated into the disagreement between observed and predicted (expressed as a ratio between observed and predicted) and the prevalence of the exposure is translated to the prevalence of the population where the disagreement exists. Therefore, PDI describes the impact of disagreement between observed and predicted risk for a risk tool as a function of the proportion of the population where that disagreement exists. PDI exemplifies how overall population risk is driven by where the cases lie in the population. This means that a large relative discrepancy between observed and predicted that is concentrated in a subgroup that covers a small proportion of the population will have less impact on the overall population estimate of diabetes compared with disagreement of the same magnitude that affects a larger proportion of the population. This finding emphasizes an important difference between individual
and population risk prediction; differences in individuals or subgroups may be important if the algorithm were to be applied to an individual, but these differences may not be as critical if applied for aggregate population estimates. This also signifies a potential difference in the way that algorithms must be validated, depending on whether they are intended for use on the individual or in small subpopulations. Of course, in the same way PAR is affected by the prevalence of the risk factor in the population, the influence of disagreement within ethnic groups is affected by the ethnic composition of the population. Ethnic composition in a population can change over time and the impact of this on the validity of the algorithm should be continually assessed. The use of prediction tools at the individual level or in small subpopulations must be independently validated in specific subpopulations and used with caution where evidence of poor fit is occurring.

The purpose of this study was to examine the impact of ethnicity on population risk prediction and not to validate it for use within specific ethnic groups. Nonetheless, looking at performance within ethnic groups provides important information about diabetes risk by ethnicity. DPoRT performed well in all ethnic groups, especially in females, with the exception of south Asian males where even with the inclusion of full ethnic information the algorithm resulted in an under-prediction of diabetes risk. This can indicate that there is an aspect of diabetes risk in this population which is not captured by either the variables in DPoRT or detailed ethnicity. This result is consistent with emerging evidence about the nature of metabolic risk in south Asian males. This population has significantly more insulin resistance than Caucasian populations even in the absence of excessive obesity (Abate and Chandalia 2001). It has been proposed that the excessive insulin resistance in Asian Indians could be explained by an abdominal fat distribution which may be genetically determined (Banerji et al. 1999). Detailed radiographic and anthropometric measures in Asian and Caucasian men showed that for a given BMI or waist circumference, south Asian men had approximately 6% higher total body fat than Caucasian men. Other studies have shown that adjustments for BMI or waist circumferences to define obesity do not entirely account for possible differences in inherent insulin resistance in the south Asian population (Chandalia et al. 1999). Several physiological mechanisms for this occurrence have been proposed including that south Asian men have a defect in adipose tissue metabolism, which occurs independently of obesity or abdominal fat distribution. These abnormalities of adipose tissue metabolism are concomitant with insulin resistance (Abate et al. 2004). These studies indicate that there may be an important aspect of diabetes risk which is not captured by simply including ethnicity and BMI along with the other predictors of the model. The type of detailed physiological information which may be needed is not captured at the population level nor is it feasible to include in a tool such as DPoRT. Regardless, these differences did not affect the performance of the model and the validity of overall population estimates of diabetes and, consistent with other prediction tools, has reaffirmed that using a measure that has a statistically significant association with a disease is not enough to improve predictive performance of a model (Pencina et al. 2008).

Another difficulty in estimating the ethnicity-diabetes risk among males is the possibility of confounding by physical activity. Immigrant men are more likely to engage in jobs that require physical activity on a daily basis (Norman et al. 2002) which has been shown to reduce the risk of developing diabetes (Qi et al. 2008). This
may explain why the full ethnicity algorithm actually performs worse than DPoRT for some ethnic groups. Inclusion of full ethnic information may result in over-fitting of the model. This phenomenon was also seen during DPoRT creation.

Previous studies indicate that weight cut-offs may differ in their associated risk for diabetes within ethnicity groups and that different cut-offs should be used to identify those at high risk (Barba et al. 2004, Diaz et al. 2007). Our study suggests that as long as additional risk factor differences among ethnic groups are captured in the prediction algorithm, the difference may not actually be as substantial as previous noted. This difference may be due to the fact that previous studies did not fully account for possible confounders including age and additional metabolic disorders (Barba et al. 2004). This study examined the interaction between age-specific BMI and ethnicity and found no significant differences.

There are several limitations to consider when interpreting the results of this study. Firstly, the minimal difference detected between DPoRT and the full ethnicity algorithm may not be found in other populations with different ethnic compositions. Secondly, using self-report survey measures is a limitation which could affect predictive risk accuracy since these measures may be more subject to reporting error and bias than clinical measures. For self-reported height and weight, in general there is a high agreement; however, validation studies show that weight tends to be slightly underestimated and that height may be slightly overestimated and as a result reported BMI is generally lower than measured BMI (Nawaz et al. 2001, Rowland 2007, Shields et al. 2008), which would result in a slight underestimation of predicted risk.

The possibility of misclassification is also possible with the use of self-reported ethnicity, even though it is the most common measure of acquiring ethnic information in epidemiological studies (Comstock et al. 2004). Interestingly self-report ethnicity is generally preferred by epidemiologists and federal agencies, such as the US and Canadian Census and the National Center for Health Statistics (Gomez et al. 2005). This is due to the fact the self-identification with ethnicity is most important for studying influences of lifestyle on disease risk. Misclassification due to self-reported ethnicity may be more problematic when examining the genetic associations with disease (Burchard et al. 2003). Physician diagnosed diabetes, as detected by claims data will not capture all diabetes cases and misclassification is possible based on the 86% specificity. Another important limitation is the exclusion of an important subpopulation which compromises the generalizability of this research to all ethnicities in Canada. The cohorts covered in the surveys used in this study exclude those living on Aboriginal reserves. Therefore, estimates for First Nations people apply only to those living off-reserve and are not intended to represent First Nations on-reserve. Previous studies show that First Nations are at greater risk for diabetes than other members of the Canadian population including off-reserve First Nations counterparts (Harris et al. 1997, Young et al. 2000, Green et al. 2003, Horn et al. 2007, Kaler et al. 2006). This is an important component of the population to consider for diabetes prevention and a population risk algorithm developed specifically for on-reserve populations would be beneficial to estimating overall diabetes burden in Canada.

This analysis provides adjusted hazard ratios and risk estimates to quantify the impact of ethnicity on diabetes risk using a prospective population-based cohort study in Ontario. This is the first study that reports 10-year risk and number of cases
of diabetes from a prediction model according to ethnicity. These estimates provide key information for predicting diabetes risk at the provincial or national level, particularly in the increasingly multiethnic Canadian population. Secondly, this study shows that DPopRT in its current form is as effective or in some cases better than the algorithm with full ethnic information for predicting diabetes risk at the population level. Furthermore, it also appears to work well within ethnic groups, in particular for women. Though overall model performance was good, analysis by ethnicity shows that further research is required to improve model fit in south Asian males.

Key messages
This study quantifies the added impact of ethnicity for predicting diabetes risk in a population-based multiethnic cohort. This is the first study that reports 10-year risk and number of cases of diabetes from a prediction model according to ethnicity. These estimates provide key information on future diabetes burden in the context of a multiethnic population. This study shows that DPopRT in its current form is as effective or in some cases better than the algorithm with full ethnic information for predicting diabetes risk at the population level. Analysis by ethnicity shows that further research is required to improve prediction in south Asian males.

Acknowledgements
This study was supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. The authors wish to thank Kelvin Lam for assistance with the creation of the figures. The study was funded by the Canadian Institutes of Health Research. 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 article. The guarantors accept full responsibility for the conduct of the study, had access to the data, and controlled the decision to publish. The study was approved by the Research Ethics Board of Sunnybrook Health Sciences Centre, Toronto, Ontario.

References
Abate, N. and Chandalia, M., 2001. Ethnicity and type 2 diabetes-Focus on Asian Indians. *Journal of Diabetes and Its Complications*, 15 (6), 320–327.
Abate, N., et al., 2004. Adipose tissue metabolites and insulin resistance in nondiabetic Asian Indian men. *Journal of Clinical Endocrinology and Metabolism*, 89 (6), 2750–2755.
Anderson, K.M., et al., 1991. An updated coronary risk profile—a statement for health-professionals. *Circulation*, 83 (1), 356–362.
Banerji, M.A., et al., 1999. Body composition, visceral fat, leptin, and insulin resistance in Asian Indian men. *Journal of Clinical Endocrinology and Metabolism*, 84 (1), 137–144.
Barba, C., et al., 2004. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet*, 363 (9403), 157–163.
Berg, A.O., et al., 2003. Screening for type 2 diabetes mellitus in adults: recommendations and rationale. *Annals of Internal Medicine*, 138 (3), 212–214.
Burchard, E.G., et al., 2003. The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine*, 348 (12), 1170–1175.
Calone, N., et al., 2008. Screening for type 2 diabetes mellitus in adults: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*, 148 (11), 846–854.

Campbell, G., 2004. General Methodology I: advances in statistic methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine*, 13, 499–508.

Canadian Diabetes Association Clinical Practice Guidelines Expert Committee, 2008. Canadian Diabetes Association 2008 clinical practice guidelines for the prevention and management of diabetes in Canada. *Canadian Journal of Diabetes*, 32 (Suppl. 1), S14–S16.

Chandalia, M., et al., 1999. Relationship between generalized and upper body obesity to insulin resistance in Asian Indian men. *Journal of Clinical Endocrinology and Metabolism*, 84 (7), 2329–2335.

Comstock, R.D., Castillo, E.M., and Lindsay, S.P., 2004. Four-year review of the use of race and ethnicity in epidemiologic and public health research. *American Journal of Epidemiology*, 159 (6), 611–619.

D’Agostino, R.B., et al., 2001. Validation of the Framingham coronary disease prediction scores. *JAMA*, 286 (2), 180–187.

Diaz, V.A., et al., 2007. How does ethnicity affect the association between obesity and diabetes? *Diabetic Medicine*, 24 (11), 1199–1204.

Eddy, D.M. and Schlessinger, L., 2003. Validation of the archimedes diabetes model. *Diabetes Care*, 26 (11), 3102–3110.

Gomez, S.L., et al., 2005. Inconsistencies between self-reported ethnicity and ethnicity recorded in a health maintenance organization. *Annals of Epidemiology*, 15 (1), 71–79.

Green, C., et al., 2003. The epidemiology of diabetes in the Manitoba-registered first nation population-Current patterns and comparative trends. *Diabetes Care*, 26 (7), 1993–1998.

Hanley, A.J.G., et al., 2003. 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*, 52 (2), 463–469.

Harrell, F.E., 2001. *Regression modeling strategies with applications to Linear Models, logistic regression, and survival analysis*. New York, NY: Springer.

Harris, S.B., et al., 1997. The prevalence of NIDDM and associated risk factors in native Canadians. *Diabetes Care*, 20 (2), 185–187.

Health Canada, 2003. *Responding to the challenge of diabetes in Canada*. Ottawa, ON: Health Canada.

Herman, W.H., et al., 1995. A new and simple questionnaire to identify people at increased risk for undiagnosed diabetes. *Diabetes Care*, 18, 382–387.

Horn, O.K., et al., 2007. Incidence and prevalence of type 2 diabetes in the first nation community of Kahnawa: ke, Quebec, Canada, 1986–2003. *Canadian Journal of Public Health-Revue Canadienne de Sante Publique*, 98 (6), 438–443.

Hux, J.E. and Ivis, F., 2005. Diabetes in Ontario. *Diabetes Care*, 25 (3), 512–516.

Ito, C., et al., 1996. Prediction of diabetes mellitus (NIDDM). *Diabetes Research and Clinical Practice*, 34 (Suppl), S7–S11.

Kaler, S.N., et al., 2006. High rates of the metabolic syndrome in a first nations community in western Canada: prevalence and determinants in adults and children. *International Journal of Circumpolar Health*, 65 (5), 389–402.

Kovacevic, M.S., Mach, L., and Roberts, G., 2008. Bootstrap variance estimation for predicted individual and population-average risks. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 2289–2296.

Levin, M.L. and Bertell, R., 1978. Re-simple estimation of population attributable risk from Case-Control Studies. *American Journal of Epidemiology*, 108 (1), 78–79.

Lindstrom, J. and Tuomilehto, J., 2007. The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care*, 26, 725–731.

Lipscombe, L.L. and Hux, J.E., 2007. Trends in diabetes prevalence, incidence, and mortality in Ontario, Canada 1995–2005: a population-based study. *Lancet*, 369 (9563), 750–756.
Lloyd-Jones, D.M., et al., 2006. Narrative review: assessment of C-reactive protein in risk prediction for cardiovascular disease. *Annals of Internal Medicine*, 145 (1), 35–42.

Mann, D.M., et al., 2010. Comparative validity of 3 diabetes mellitus risk prediction scoring models in a multiethnic US cohort. *American Journal of Epidemiology*, 171 (9), 980–988.

Manuel, D. and Schultz, S., 2003. *Diabetes in Ontario: an ICES practice atlas*. Toronto, ON: Institute for Clinical and Evaluative Sciences.

Mokdad, A.H., et al., 2001. The continuing epidemics of obesity and diabetes in the United States. *JAMA: The Journal of the American Medical Association*, 286 (10), 1195–1200.

Nam, B.-H., 2000. Discrimination and calibration in survival analysis: extension of the ROC curve for discrimination and chi-square test for calibration. Boston, New England: Boston University.

Nawaz, H., et al., 2001. Self-reported weight and height: implications for obesity research. *Journal of Preventive Medicine*, 20, 294–298.

Newbold, K.B. and Danforth, J., 2003. Health status and Canada’s immigrant population. *Social Science & Medicine*, 57 (10), 1981–1995.

Norman, A., et al., 2002. Total physical activity in relation to age, body mass, health and other factors in a cohort of Swedish men. *International Journal of Obesity*, 26 (5), 670–675.

Norris, S.L., et al., 2008. Screening adults for type 2 diabetes: a review of the evidence for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, 148 (11), 855–868.

Odea, K., 1991. Westernization, insulin resistance and diabetes in Australian Aborigines. *Medical Journal of Australia*, 155 (4), 258–264.

Odea, K., et al., 1993. Obesity, diabetes, and hyperlipidemia in a Central Australian Aboriginal Community with a long history of acculturation. *Diabetes Care*, 16 (7), 1004–1010.

Odell, P.M., Anderson, K.M., and Kannel, W.B., 1994. New models for predicting cardiovascular events. *Journal of Clinical Epidemiology*, 47 (6), 583–592.

Pavkov, M.E., et al., 2007. Changing patterns of type 2 diabetes incidence among Pima Indians. *Diabetes Care*, 30 (7), 1758–1763.

Pencina, M. and D’Agostino, R.B., 2004. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23, 2109–2123.

Pencina, M.J., et al., 2008. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27 (2), 157–172.

Pepe, M.S., et al., 2004. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, 159, 882–890.

Qi, L., Hu, F.B., and Hu, G., 2008. Genes, environment, and interactions in prevention of type 2 diabetes: a focus on physical activity and lifestyle changes. *Current Molecular Medicine*, 8 (6), 519–532.

Ramachandran, A., et al., 1997a. Rising prevalence of NIDDM in an urban population in India. *Diabetologia*, 40 (2), 232–237.

Ramachandran, A., et al., 1997b. Risk of noninsulin dependent diabetes mellitus conferred by obesity and central adiposity in different ethnic groups: a comparative analysis between Asian Indians, Mexican Americans and Whites. *Diabetes Research and Clinical Practice*, 36 (2), 121–125.

Rosella, L.C., et al., 2010. A population based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). *Journal of Epidemiology and Community Health*, doi:10.1136/jech.2009.102244

Rowland, M., 2007. Self-reported height and weight. *American Journal of Clinical Nutrition*, 52, 1125–1133.

Shields, M., Gorber, S.C., and Tremblay, M.S., 2008. Estimates of obesity based on self-report versus direct measures. *Health Reports*, 19 (2), 1–16.

Statistics Canada, 1999. *1996–7 National Population Health Survey: derived variable specifications*. Ottawa, ON: Statistics Canada.

Statistics Canada, 2001. *Population and dwelling counts, for Census Divisions, Census Subdivisions (Municipalities) and designated places, 2001 and 1996*. Ottawa, ON: Statistics Canada.
Statistics Canada, 2002. Canadian Community Health Survey Methodological Overview. Health Reports, 13, 9–14.
Statistics Canada, 2003. Canadian Community Health Survey, 2000–2001. Ottawa, ON: Statistics Canada.
Wild, S., et al., 2004. Global prevalence of diabetes-estimates for the year 2000 and projections for 2030. Diabetes Care, 27, 1047–1053.
Wilson, P.W.F., et al., 2005. C-reactive protein and risk of cardiovascular disease in men and women from the Framingham Heart Study. Archives of Internal Medicine, 165 (21), 2473–2478.
World Health Organization, 1998. Report of a WHO consultation on obesity, Obesity: preventing and managing the global epidemic. Geneva: World Health Organization.
Yeo, D., Mantel, H., and Lui, T.P., 1999. Bootstrap variance estimation for the National Population Health Survey. Baltimore: American Statistical Association.
Young, T.K., et al., 2000. Type 2 diabetes mellitus in Canada’s First Nations: status of an epidemic in progress. Canadian Medical Association Journal, 163 (5), 561–566.