BMJ Public Health

Development and validation of a population-based risk algorithm for premature mortality in Canada: the Premature Mortality Population Risk Tool (PreMPoRT)

Meghan O'Neill,[1] Mackenzie Hurst,[1] Lief Pagalan ,[1] Lori Diemert,[1] Kathy Kornas,[1] Stacey Fisher,[1] Andy Hong,[2] Doug Manuel ,[3] Laura C Rosella [1]

To cite: O'Neill M, Hurst M, Pagalan L, *et al*. Development and validation of a population-based risk algorithm for premature mortality in Canada: the Premature Mortality Population Risk Tool (PreMPoRT). *BMJ Public Health* 2024;**2**:e000377. doi:10.1136/bmjph-2023-000377

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjph-2023-000377).

Received 4 July 2023
Accepted 14 August 2024

[1]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[2]Department of City & Metropolitan Planning, University of Oxford, Salt Lake City, Utah, USA
[3]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

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

ABSTRACT

Introduction To develop and validate the Premature Mortality Population Risk Tool (PreMPoRT), a population-based risk algorithm that predicts the 5-year incidence of premature mortality among the Canadian adult population.

Methods Retrospective cohort analysis used six cycles of the Canadian Community Health Survey linked to the Canadian Vital Statistics Database (2000–2017). The cohort comprised 500 870 adults (18–74 years). Predictors included sociodemographic factors, self-perceived measures, health behaviours and chronic conditions. Three models (minimal, primary and full) were developed. PreMPoRT was internally validated using a split set approach and externally validated across three hold-out cycles. Performance was assessed based on predictive accuracy, discrimination and calibration.

Results The cohort included 267 460 females and 233 410 males. Premature deaths occurred in 1.40% of females and 2.05% of males. Primary models had 12 predictors (females) and 13 predictors (males). Shared predictors included age, income quintile, education, self-perceived health, smoking, emphysema/chronic obstructive pulmonary disease, heart disease, diabetes, cancer and stroke. Male-specific predictors were marital status, Alzheimer's disease and arthritis while female-specific predictors were body mass index and physical activity. External validation cohort differed slightly in demographics. Female model performance: split set (c-statistic: 0.852), external (c-statistic: 0.856). Male model performance: split set and external (c-statistic: 0.846). Calibration showed slight overprediction for high-risk individuals and good calibration in key subgroups.

Conclusions PreMPoRT achieved the strongest discrimination and calibration among existing prediction models for premature mortality. The model produces reliable estimates of future incidence of premature mortality and may be used to identify subgroups who may benefit from public health interventions.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Premature mortality is a significant population health issue, often preventable through policy, preventive interventions or treatments. Existing prediction models for premature mortality do not use widely available population-based survey data and demonstrate moderate accuracy and calibration.

WHAT THIS STUDY ADDS

⇒ We developed and validated the Premature Mortality Population Risk Tool (PreMPoRT), which demonstrates strong reproducibility and transportability. It achieves the highest discrimination and calibration among existing prediction models for premature mortality using routinely collected data.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ PreMPoRT provides a robust tool for public health applications to predict the risk of premature mortality using widely available survey data. This model can support targeted public health interventions and policy decisions to reduce premature mortality, enhancing the precision of health promotion and disease prevention strategies at the population level.

INTRODUCTION

Premature mortality is an important health outcome worldwide as it reflects an unfulfilled life expectancy and is indicative of population health, socioeconomic disparities and the effectiveness of public health interventions.[1] It is suggested that over 40% of premature deaths are avoidable through policy, prevention or treatment.[2–4] Premature mortality is frequently used as an international comparator for tracking inequalities and has been used as a benchmark to assess overall health system functioning.[5–7] For example,

reductions in premature mortality were recently highlighted as a target set by the United Nations Sustainable Development Goals.[5] While generally, life expectancy is projected to increase in high-income nations, some countries have seen stalls or even increases in premature mortality in recent years.[8 9] Additionally, trends suggest that the socioeconomic inequalities between those in the lowest and highest socioeconomic strata are growing.[10–13]

Increasingly, disease and mortality risk information is used to improve how health interventions are designed. For populations, the concept of using risk information to better target interventions has recently been described as precision public health. This approach involves the use of rich population-based data to inform which subgroups should receive targeted interventions at the optimal time.[14] One tool that enables such data decisions is prediction models.[15] Population prediction models leverage population-based data to determine individual characteristics to predict an outcome.[16] Such models have predominantly been deployed in clinical settings, typically with data from electronic health records to estimate an individual's future risk of a health outcome.[17] With increases in data availability at the population level, the ability to leverage prediction models for public health planning has grown.[15 18] Prediction models afford several advantages for informing resource allocation and populations in need of targeted interventions in an applied public health setting as they allow the user to predict risk with only baseline exposure data (ie, users of the prediction model do not need linked mortality data).

To the best of our knowledge, there is no population-based risk prediction tool for estimating the future incidence of premature mortality in the Canadian adult population. To date, there have been few mortality prediction models for population and public health applications and even fewer for premature mortality. There have only been two studies that focused on prediction for all-cause mortality at a population level. One study by Manuel *et al* used the same data source as our study to predict all-cause mortality[19] and another used data from the UK Biobank study to predict all-cause mortality in the UK population.[20] To date, there has only been one published prediction model for premature mortality, which was developed by Weng *et al*, using UK Biobank data. This study differs from ours for two important reasons. First, the study aim appears to be more comparative in nature (ie, the comparison of machine learning methods to traditional survival approaches for prediction) and is different from ours in the type of data used. For example, we use self-reported population-based survey data while the study by Weng *et al* incorporates clinical predictors such as medication history and biometric measurements.[21] In this study, we developed and validated the Premature Mortality Population Risk Tool (PreMPoRT), a population-based risk algorithm that predicts the 5-year incidence of premature mortality among the Canadian adult population.

METHODS

Before receiving the data, we published an analytical protocol to increase the robustness of our approach and transparency in the model development and validation processes.[22] We describe in extensive detail the prespecified model predictors, statistical analysis plan, sample size calculation, coding and cleaning of data, approach to missing data, model estimation, model specification, model validation and assessment of model performance.[22] These steps are also described in brief below. Overall, the analytical protocol was adhered to with some modifications, which we describe. First, an additional cohort exclusion included individuals residing in the Yukon, Northwest Territories and Nunavut because area-based information and household income are unavailable for those regions. Second, we considered two additional predictor variables for inclusion in our model which were not initially specified, specifically food insecurity and area-level income. Third, in addition to the primary model specified in the protocol, a minimal and full model was developed and validated. Fourth, the minimal model was fit with the fewest variables to allow for user flexibility when all model variables may not be available. The maximal model where more variables were fitted was generated for comparison purposes. Fifth, internal validation consisted of a split-set instead of the bootstrap approach as a more robust form of internal validation and to eliminate issues with model stability using the bootstrap approach. We adhered to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guidelines, which are the best practice standards for the development and validation of prediction models.[23]

Data sources

A retrospective cohort study was used to develop a prediction model for premature mortality using population-based survey data from the Canadian Community Health Survey (CCHS). The CCHS is a cross-sectional survey covering 98% of the Canadian population 12 years and older that collects self-reported data on personal health status, healthcare utilisation and health determinants.[24] Populations excluded from the CCHS sampling frame include individuals living in Aboriginal settlements, Canadian Force Bases and some remote regions. A detailed description of the CCHS survey methodology is available elsewhere.[24] All respondents were linked to the Canadian Vital Statistics Database (CVSD) to ascertain deaths during the follow-up period. The data were held at the Statistics Canada Research Data Centre. Due to Research Data Centre vetting requirements, all study output is rounded deterministically to the nearest 10.

Participants

Respondents who consented to link their responses to administrative data from the first six cycles 1.1 (2000/2001), 2.1 (2003/2004), 3.1 (2005/2006), 2007/2008, 2009/2010 and 2011/2012 of the CCHS

were used to create the study cohort. Respondents were excluded if they were pregnant, resided in the territories (ie, Yukon, Northwest Territories or Nunavut) or were under 18 years or older than 74 years at the date of CCHS interview. For further details on our cohort creation, please see online supplemental figure S1.

Outcome: premature mortality

Respondents were followed longitudinally for the incidence of premature mortality (binary yes/no variable), defined as deaths from all causes between the ages of 18 and 74 as recorded in the CVSD. This definition is based on the cut-off adopted by the Canadian Institute of Health Information,[25] which is consistent with the definition used in reporting premature mortality among other industrialised nations.[26–29]

Predictors

As stated in our protocol, 38 candidate predictors were identified from the survey data in accordance with established associations with premature mortality,[20 30–34] subject matter knowledge, user input, our team's experience with the development and validation of population-level risk algorithms[19 35–41] and predictor availability across survey cycles. Predictors included sociodemographic characteristics, self-perceived measures, health behaviours, chronic conditions and area-based measures.

Missing data

Multiple imputation methods were used to impute missing data, specifically, Fully conditional specification (FCS) was used to develop five imputed datasets which previous research has established as sufficient.[42 43] Total missingness was low and ranged in the six combined cycles from <1% to 10%.[22] All imputation was done using the multivariate imputation by chained equations (mice) algorithm in R.[42] First, each cycle was separated into sex-stratified groups and imputation was run on each sex-stratified cycle separately to avoid potential between cycle variations and differences between males and females. A three-step approach was used to impute the missing values. FCS imputation was unable to converge for several chronic conditions that had a low prevalence and low missing data (ie, less than 1% missing). As such, the first step was to assume that a respondent did not have said chronic condition if any of the chronic conditions had less than 1% missing across all sex-stratified cycles. These chronic conditions included Alzheimer's disease, arthritis, asthma, back problems, bowel disease, cancer, chronic obstructive pulmonary disease (COPD)/emphysema, diabetes, high blood pressure, heart disease, intestinal ulcers, migraines, stroke and urinary incontinence. Afterwards, five imputed datasets were generated using FCS on all sociodemographics, health behaviours and chronic conditions except for anxiety and mood disorders. Anxiety and mood disorders were not asked in the CCHS cycle 1.1, and as such, a third and final step needed to be implemented. Each imputed dataset was divided by sex and by cohort group (ie, derivation cohort and validation cohort). Once separated, FCS was run to impute the missing values for anxiety and mood disorders.

A sensitivity analysis was conducted to examine the variation in model performance measures across the imputed datasets, which would help inform whether we averaged the measures or reported performance measures on the first imputed dataset only. After finding negligible differences in the model performance measures (online supplemental table S1), we opted to present model development and validation measures on the first imputed dataset.

Study design

Models were developed and validated in the Canadian adult provincial population.[44 45] There were five steps involved in the development and validation: (1) model derivation in the first three CCHS cycles ((1.1 (2000/2001), 2.1 (2003/2004) and 3.1 (2005/2006)). Two models, one for females and one for males, were developed due to the important biological sex differences in premature mortality[34 46] (further outlined in the 'Model specification' section). This was followed by (2) the internal validation using a split sample approach where the 70% development model was applied to the remaining 30%. The model was then (3) externally validated in the last three CCHS cycles (2007/2008, 2009/2010 and 2011/2012). Next, (4) the derivation and validation data were combined to estimate the final application of the PreMPoRT model (ie, the final Canadian provincial cohort) using the same model specification as in the original model derivation. Lastly, (5) the predictive performance of the final primary model was assessed among more than 20 programme and policy-relevant subgroups with <20% difference between observed and predicted, representing good calibration.[15 19] In the main results, we report observed and predicted risk for three of these subgroups (ie, education, income and immigration status).

Model specification

We began with the prespecified forms of predictors as described in the study protocol (online supplemental table S2). Candidate predictors that did not improve predictive performance were removed. Next, the primary model was selected from the minimal model, using the sequential addition of predictors until the performance resembled that of the full model. Finally, we verified the sequential addition of predictors in each of the three models using the least absolute shrinkage and selection operator as described in our study protocol.[22] Following sequential predictor selection, alternative specifications were examined. For example, both BMI and age were respecified from continuous with restricted cubic splines to categories for both sexes. The focus of the results is on the primary model, however, results for the minimal and full models are in online supplemental file 1.

Statistical analysis

The predictive performance was based on measures of overall predictive accuracy (ie, Nagelkerke R^2, Brier score), discrimination (ie, Harrell's c-statistic, time-specific discrimination slope) and calibration (ie, time-specific calibration curve, calibration intercept and calibration slope). The definitions of each measure are found in online supplemental table S3. From January 2000 to December 2017, respondents from the CCHS were followed until premature mortality, 5 years of follow-up or study end date. The models were estimated using a Weibull accelerated failure time model with the proportional hazard's assumption assessed according to the scaled Schoenfeld residuals. The CCHS survey weights were applied to ensure all estimates were representative of the population.[47 48] Data cleaning and predictor coding were conducted in SAS V.9.4, with all model development and validation executed in R V.3.6.2.

Patient and public involvement

This study and the published analytical protocol were developed in consultation with decision-makers at three local public health departments in urban and rural Ontario to ensure that the design of the prediction model is optimised for applied public health applications. These decision-makers were engaged at the beginning of this work to provide feedback on the published study protocol which detailed the design, analysis and reporting of the present study.

RESULTS

Study population characteristics

Baseline characteristics of the entire Canadian female provincial cohort (n=267 460), male provincial cohort (n=233 410), and external development and validation cohorts (which are a subset of Canadian provincial cohort) are detailed in table 1. The average number of premature deaths over the study period was 1.40% for females and 2.05% for males. Among the Canadian provincial cohort, female respondents were 44.18 years and 43.43 years for males at survey date. Most individuals had completed post-secondary education, and the most common marital status was domestic partnership (married/common law). The most common BMI category for females was 18.5 to <25.0 kg/m^2, whereas for males, it was 25.0 to <30.0 kg/m^2. Most participants rated their self-perceived general health as good or very good. Physical activity levels were lower in females than in males, with 51.08% of females classified as inactive compared with 46.41% of males. Most participants were white, Canadian-born and lived in urban areas. Differences in the population characteristics were greatest in the external development and validation cohort compared with the split set development and validation cohort (online supplemental table S4). For example, there were larger proportions in the lowest income quintile, larger proportions of people with post-secondary education, more never-smokers and greater representation of visible minorities.

Model development

The full female and male model consisted of 16 predictors, the primary female model consisted of 12 predictors and the male model consisted of 13 predictors, the minimal female model consisted of 8 predictors and the male model contained 10 predictors. The model coefficients and HRs for the primary model are outlined in table 2. Age, self-rated health, smoking, COPD/emphysema, heart disease, diabetes, cancer and stroke were the predictors with the largest contributions to model performance and appeared in all versions (ie, minimal, primary and full) of the sex-specific models. We assessed the proportionality assumption (via scaled Schoenfeld residuals) inherent in Weibull accelerated failure time models and found no violations. The model coefficients and HRs for the minimal and full models are found in online supplemental table S5 (females) and online supplemental table S6 (males).

Model external development

The predictive performance of the primary female and male models in different development and validation settings are described in table 3. The primary development cohort was the external development (CCHS cycles 2000–2006). In the primary development cohort, the female model achieved better predictive performance than the male model. The female external development model achieved a Brier score: 0.007, discrimination slope: 0.096 and c-statistic: 0.895. The male external development model also performed well (Brier score: 0.010) and had good discriminative ability (discrimination slope: 0.105 and c-statistic: 0.853).

Model split sample and external validation

The female model in the external validation cohort produced the best overall predictive performance (Brier score: 0.006) and discrimination (discrimination slope: 0.109) (table 3). The female model also performed well in the split set validation (Brier score: 0.007 and discrimination slope: 0.094). The male model performed slightly better in the split set validation in terms of predictive performance (Brier score: 0.010) and discrimination (discrimination slope: 0.096) in comparison to the external validation (Brier score: 0.009 and discrimination slope: 0.108). The c-statistic for the female split set validation was 0.852 and the male was 0.846 (table 3). The female and male split set calibration curves (figures 1A,C) exhibit similar shapes, with strong adherence to the 45° line for low levels of predicted risk and a tendency to overpredict risk among the small proportion of high-risk individuals. The c-statistic for the female and males models in the external validation cohorts was 0.856 and 0.846, respectively (table 3). The calibration

Table 1 Weighted baseline characteristics of the entire Canadian provincial cohort and external development and validation cohorts, by sex

Characteristic	Canadian provincial cohort (CCHS cycles 2000–2012)		External development cohort (2000–2006)		External validation cohort (CCHS cycles 2007–2012)	
	Female (n=267 460)	Male (n=233 410)	Female (n=138 070)	Male (n=121 890)	Female (n=129 390)	Male (n=111 520)
Premature deaths (%)	1.40	2.05	1.42	2.09	1.39	2.04
Follow-up time, mean (SD), years	4.78 (0.83)	4.78 (0.82)	4.78 (0.82)	4.79 (0.81)	4.77 (0.84)	4.77 (0.84)
Age, mean (SD), years	44.18 (0.05)	43.43 (0.06)	43.49 (0.06)	42.73 (0.06)	44.50 (0.074)	43.76 (0.078)
Age group						
18–24	12.70	13.58	12.94	13.91	12.59	13.42
25–34	17.41	18.49	17.62	18.50	17.32	18.48
35–44	20.58	20.87	23.15	23.53	19.40	19.64
45–54	21.26	20.45	20.87	20.18	21.44	20.57
55–64	16.97	16.78	14.91	14.61	17.91	17.78
65–74	11.08	9.85	10.51	9.27	11.34	10.11
Household income quintile						
Q1 (lowest)	17.34	12.94	9.68	7.31	20.87	15.53
Q2	17.77	15.01	12.10	9.13	20.39	17.73
Q3	20.70	19.70	21.31	18.82	20.42	20.10
Q4	22.95	25.29	30.40	31.22	19.51	22.55
Q5 (highest)	21.24	27.06	26.51	33.52	18.82	24.09
Individual education						
Less than secondary school graduation	7.05	6.36	9.19	8.09	6.07	5.56
Secondary school graduation	10.80	10.93	11.97	11.82	10.26	10.51
Post-secondary education (complete and partial)	82.15	82.72	78.84	80.09	83.67	83.93
Marital status						
Single never married	22.36	26.91	21.37	26.45	22.81	27.12
Domestic partner (married/common law)	63.39	65.54	64.45	66.45	62.90	65.12
Widowed/separated/divorced	14.25	7.55	14.18	7.10	14.29	7.76
BMI, mean (SD), (kg/m^2)	25.38 (0.02)	26.57 (0.02)	25.07 (0.02)	26.32 (0.02)	25.52 (0.03)	26.69 (0.03)
BMI categories						
<18.5	3.93	1.12	4.31	1.15	3.76	1.10
18.5 to <25.0	52.29	39.70	54.04	41.53	51.48	38.85
25.0 to <30.0	26.93	40.68	26.55	40.69	27.10	40.68
30.0 to <35.0	11.13	14.10	10.26	40.69	11.54	14.61
35.0 to <40.0	3.75	3.23	3.25	2.73	3.98	3.46
≥40.0	1.97	1.17	1.60	0.89	2.14	1.30
Self-perceived general health						
Poor	2.72	2.44	2.76	2.41	2.71	2.45
Fair	8.54	7.83	8.68	7.71	8.48	7.89
Good	28.28	28.56	28.31	28.15	28.26	28.75
Very good	38.21	38.12	37.33	37.11	38.61	38.59
Excellent	22.26	23.05	22.93	24.62	21.95	22.32
Physical activity*						
Active	23.15	28.47	21.48	26.79	23.92	29.24
Moderately active	25.77	25.12	25.53	24.82	25.88	25.27

Continued

Table 1 Continued

Characteristic	Canadian provincial cohort (CCHS cycles 2000–2012)		External development cohort (2000–2006)		External validation cohort (CCHS cycles 2007–2012)	
	Female (n=267 460)	Male (n=233 410)	Female (n=138 070)	Male (n=121 890)	Female (n=129 390)	Male (n=111 520)
Inactive	51.08	46.41	52.99	48.39	50.20	45.49
Smoking status						
Never-smoker	56.37	46.69	53.55	43.69	57.66	48.08
Current light smoker	18.56	20.84	20.18	21.43	17.82	20.57
Current heavy smoker	2.74	5.59	3.83	6.86	2.24	5.00
Former light smoker	17.21	17.37	16.85	17.15	17.38	17.47
Former heavy smoker	5.11	9.51	5.59	10.87	4.89	8.88
Self-reported chronic conditions						
Arthritis	18.40	11.59	19.56	12.03	17.86	11.39
Emphysema/COPD	1.94	1.62	0.85	0.98	2.45	1.92
Heart disease	3.26	4.84	3.37	4.85	3.21	4.83
Diabetes	4.88	6.02	4.08	4.90	5.25	6.54
Cancer	1.91	1.55	1.63	1.37	2.04	1.64
Stroke	0.84	0.87	0.74	0.86	0.89	0.88
Bowel disease	5.87	2.43	4.18	1.75	6.65	2.75
Ethnicity†						
Visible minority	16.75	16.55	15.16	15.17	17.49	17.19
White	83.25	83.45	84.84	84.83	82.51	82.81
Immigration status						
Canadian born	77.14	77.29	78.85	78.78	76.35	76.61
Immigrant (<10 years)	6.53	6.57	5.83	5.78	6.85	6.93
Immigrant (≥10 years)	16.34	16.14	15.32	15.43	16.81	16.46
Rurality						
Rural	17.55	18.58	17.83	18.92	17.42	18.43
Urban	82.45	81.42	82.17	81.08	82.58	81.57

*Physical activity was measured using the average metabolic equivalent of task per day derived from a list of leisure-time physical activities (frequency and duration of activity).
†The CCHS collects data on ethnicity by asking respondents to identify if they belong to a visible minority group. This terminology represents that employed by Statistics Canada.[56]
BMI, body mass index; CCHS, Canadian Community Health Survey; COPD, chronic obstructive pulmonary disease.

curves for the female and male models in external validation (figure 1B,D) diverge from the 45° line at a lower predicted risk than in the split set validation. However, given that <2% of the overall cohort has a predicted risk above 20%, there was minimal impact of the overprediction on overall model calibration. Calibration curves and predictive performance measures for the minimal and full model in the split set and external validation are found in online supplemental figures S3 and S4 and table S7, respectively.

Subgroup calibration

The model was further evaluated on more than 20 important policy or programme-relevant population subgroups. In figure 2, we report the observed incidence of premature mortality compared with average model prediction among females and males for three of the 20 subgroups we assessed, including education, income and immigration status. Overall, calibration across subgroups was strong. Among both sexes, we observed minor underprediction among those with less secondary school graduation and post-secondary education. Among those with less than secondary school graduation, there is slight overprediction for males and underprediction for females. With the exception of income quintile 2 for males, there is slight underprediction for both sexes across income quintiles. Among both sexes, there was minor overprediction in immigrants and visible minorities and slight underprediction for people born in Canada.

Table 2 The proportional hazards coefficients and HRs for the primary female and male model in the Canadian Provincial Cohort

	Female		Male	
Variable	**ln (HR) (95% CI)**	**HR**	**ln (HR) (95% CI)**	**HR**
Weibull parameters				
Scale parameter	0.819		0.847	
Shape parameter	1.221		1.181	
Intercept	−6.956 (−6.981 to −6.932)	0.001	−6.341 (−6.359 to −6.323)	0.002
Age	0.076 (0.075 to 0.076)	1.079	0.075 (0.074 to 0.075)	1.078
Household income quintile				
Q1 (lowest)	0.225 (0.207 to 0.243)	1.252	0.328 (0.314 to 0.342)	1.388
Q2	0.143 (0.125 to 0.161)	1.153	0.382 (0.368 to 0.396)	1.465
Q3	0.057 (0.039 to 0.075)	1.058	0.240 (0.226 to 0.253)	1.271
Q4	0.086 (0.068 to 0.103)	1.089	0.114 (0.100 to 0.127)	1.121
Q5 (highest)	Ref.	Ref.	Ref.	Ref.
Marital status				
Single never married	NA	NA	Ref.	Ref.
Domestic partner (married/common law)	NA	NA	−0.564 (−0.576 to −0.551)	0.569
Widowed/separated/divorced	NA	NA	−0.094 (−0.108 to −0.080)	0.911
Education				
Less than secondary school graduation	Ref.	Ref.	Ref.	Ref.
Secondary school graduation	0.092 (0.078 to 0.105)	1.096	0.138 (0.127 to 0.149)	1.148
Post-secondary education (complete and partial)	0.013 (−0.001 to 0.028)	1.013	0.075 (0.063 to 0.087)	1.078
Body mass index (kg/m^2)				
<18.5	0.424 (0.402 to 0.446)	1.528	NA	NA
18.5 to <25.0	Ref.	Ref.	NA	NA
25.0 to <30.0	−0.297 (−0.309 to −0.285)	0.743	NA	NA
30.0 to <35.0	−0.411 (−0.427 to −0.395)	0.663	NA	NA
35.0 to <40.0	−0.373 (−0.396 to −0.350)	0.689	NA	NA
≥40.0	−0.167 (−0.194 to −0.139)	0.846	NA	NA
Self-perceived general health				
Poor	1.847 (1.823 to 1.871	6.340	1.689 (1.670 to 1.708)	5.414
Fair	1.043 (1.021 to 1.065	2.837	0.903 (0.886 to 0.920)	2.468
Good	0.541 (0.521 to 0.561)	1.717	0.462 (0.447 to 0.478)	1.588
Very good	0.170 (0.149 to 0.190)	1.185	0.116 (0.100 to 0.131)	1.122
Excellent	Ref.	Ref.	Ref.	Ref.
Cigarette smoking				
Never smoked	Ref.	Ref.	Ref.	Ref.
Former light smoker	0.475 (0.461 to 0.489	1.608	0.196 (0.183 to 0.209)	1.216
Former heavy smoker	0.676 (0.489 to 0.694	1.967	0.469 (0.456 to 0.481)	1.598
Current light smoker	0.865 (0.851 to 0.879)	2.375	0.732 (0.719 to 0.744)	2.079
Current heavy smoker	1.015 (0.995 to 1.035)	2.760	1.058 (1.043 to 1.072)	2.879
Physical activity				
Inactive	0.332 (0.317 to 0.348)	1.394	NA	NA
Moderately active	0.107 (0.089 to 0.125)	1.113	NA	NA
Active	Ref.	Ref.	NA	NA
Chronic conditions				
Emphysema/COPD	0.188 (0.170 to 0.206)	1.207	0.209 (0.194 to 0.225)	1.233

Continued

Table 2 Continued

Variable	Female		Male	
	ln (HR) (95% CI)	HR	ln (HR) (95% CI)	HR
Heart disease	0.188 (0.173 to 0.203)	1.206	0.224 (0.213 to 0.235)	1.251
Diabetes	0.271 (0.256 to 0.285)	1.311	0.285 (0.275 to 0.296)	1.330
Cancer	1.667 (1.653 to 1.682)	5.298	1.246 (1.233 to 1.259)	3.477
Stroke	0.325 (0.303 to 0.348)	1.385	0.236 (0.217 to 0.254)	1.266
Alzheimer's disease	NA	NA	0.692 (0.658 to 0.726)	1.998
Arthritis	NA	NA	−0.285 (−0.295 to −0.275)	0.752

COPD, chronic obstructive pulmonary disease; NA, not available.

DISCUSSION

Principal findings

We developed and externally validated a prediction model for population health applications to determine individual risk of premature mortality within 5 years. Model predictors from routinely collected survey data included sociodemographic characteristics, self-perceived measures, health behaviours and chronic conditions. The most influential predictors, with the largest contributions to model performance, included age, self-rated health, smoking, COPD/emphysema, heart disease, diabetes, cancer and stroke. The final model showed reproducibility and transportability through different validation settings (ie, internal and external validation). Further, PreMPoRT demonstrated robust discrimination and calibration across different subpopulations. Thus, we anticipate that this tool will produce valid and accurate predictions of premature mortality when applied to support interventions and health policies aimed at reducing premature mortality in the population.

Strengths and limitations

One of the main strengths of this study is the ability to validate a robust prediction model, based on routinely collected national population-based survey data that is representative of the national population. This dramatically increases the ability for this model to be used widely and regularly, compared with many existing models that rely on the use of highly restricted data. Due to the use of a national mortality registry, we were able to comprehensively capture all premature deaths over the study period. In addition, to reduce the possibility of overfitting and to improve transparency in the development of PreMPoRT we prespecified our analytical plan in our study protocol.[22] Beyond use within Canada's population and public health system, countries internationally

Table 3 Measures of predictive performance for the primary female and male models, by development and validation setting*

Measures of predictive performance	Final Canadian provincial model	Split set validation		External validation	
		Development (70%)	Validation (30%)	Development (2000–2006)	Validation (2007–2012)
Female model					
Nagelkerke's R^2	0.193	0.197	NA	0.192	NA
Brier score	0.007	0.006	0.007	0.007	0.006
Discrimination slope	0.094	0.096	0.094	0.096	0.109
Harrell's c-statistic	0.858	0.859	0.852	0.859	0.856
Calibration in the large	−0.007	−0.007	−0.007	−0.006	−0.010
Calibration slope	1.000	1.000	0.963	1.0000	0.930
Male model					
Nagelkerke's R^2	0.184	0.185	NA	0.190	NA
Brier score	0.009	0.009	0.010	0.010	0.009
Discrimination slope	0.099	0.099	0.096	0.105	0.108
Harrell's c-statistic	0.849	0.849	0.846	0.853	0.846
Calibration in the large	−0.012	−0.012	−0.012	−0.011	−0.018
Calibration slope	1.000	1.000	0.997	1.000	0.970

*Definitions of model performance measures can be found in online supplemental table S2.
NA, not available.

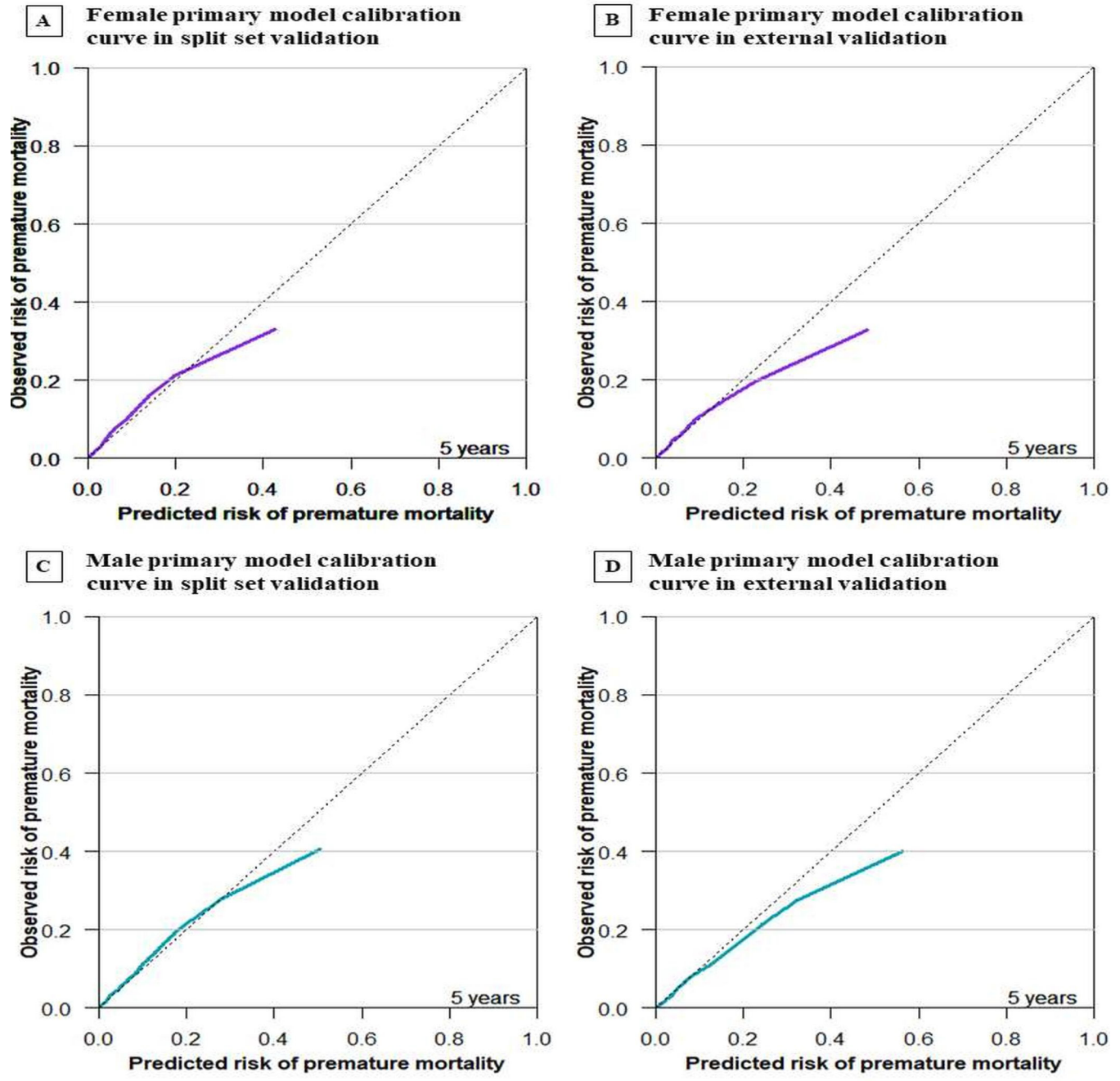

Figure 1 Female and male calibration curves for the primary model in the internal split set and external validation setting.

are routinely employing population health surveys with increasing standardisation.[49] Given that PreMPoRT was developed using data from the Canadian population, findings are likely generalisable to other high-income nations with similar population profiles. However, it is important to consider other contextual factors, including variations in healthcare systems, social determinants of health and health status, which may influence the performance and applicability of the tool in different countries. We recommend users update the baseline hazard function before employing this tool in other countries. Lastly, this study is enhanced by the adherence to best practices and reported in line with TRIPOD guidelines.[23]

This prognostic study should be interpreted considering a few limitations. Although the CCHS is representative of 98% of the population, due to the nature of the sampling methodology the data does not include people living in correctional facilities, long-term care and retirement homes, and those living in Aboriginal settlements.[24] Although these groups represent <3% of the CCHS target population, it is important given that these populations have different premature mortality risk than the general population.[50 51] Additionally, baseline risk factors for the development and validation of PreMPoRT were collected through the CCHS, which is a cross-sectional, self-reported survey. As a result, there is potential for

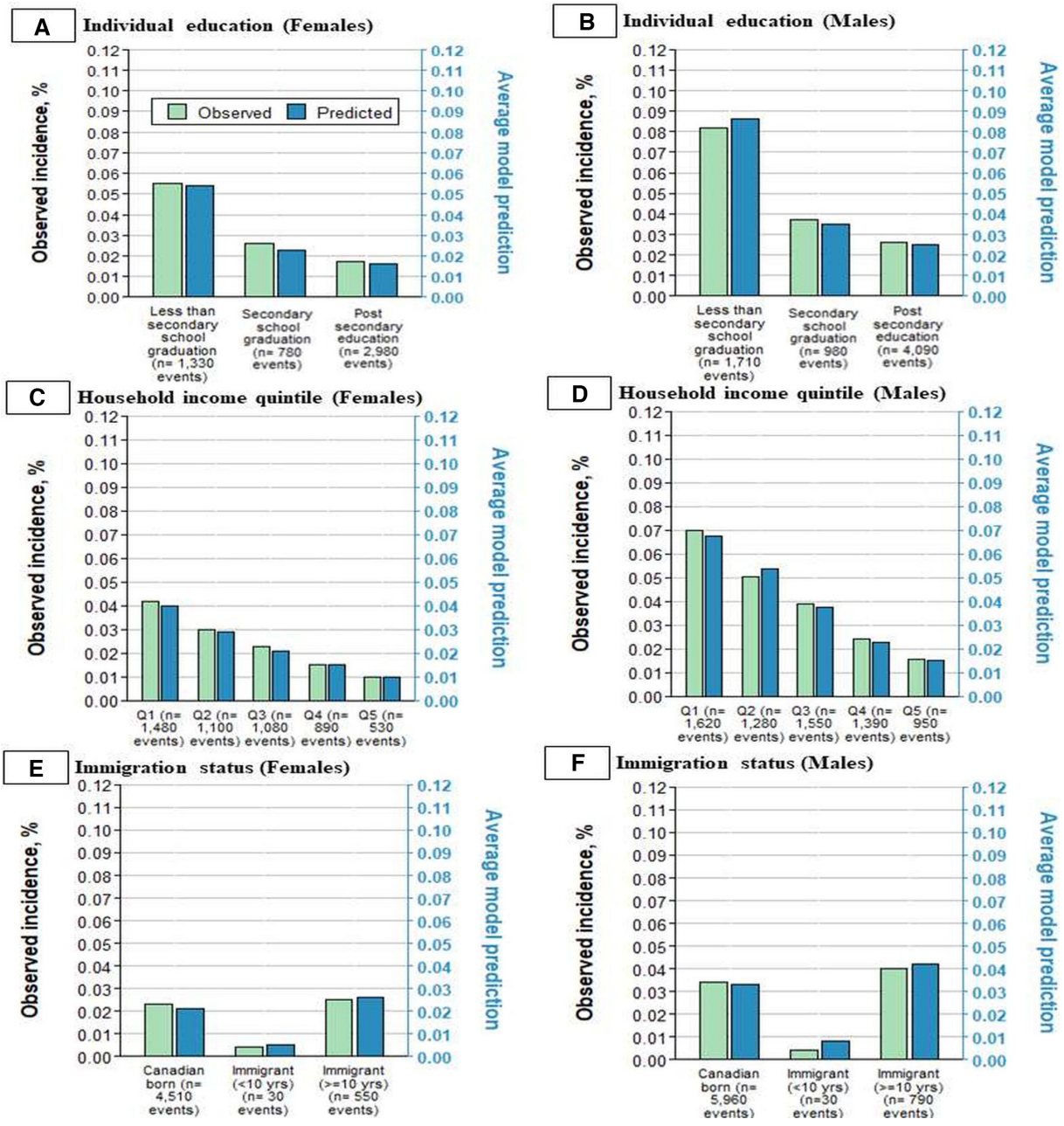

Figure 2 Premature mortality risk among female and male population subgroups.

misclassification (systematic and non-directional). Finally, miscalibration was present in <2% of the population with the highest predicted risk. In these populations, public health decision-makers may interpret the precise probabilities with caution but should still treat these individuals as higher risk than the general population. Additionally, this minor overprediction may be corrected with recalibration techniques.

Comparison with previous literature

To the best of our knowledge, there are no systematic reviews that capture development and validation studies of prediction models for premature mortality. There is one study that describes the development and validation of a prediction model for premature mortality by Weng *et al*, that compared the use of traditional survival methods to machine learning approaches.[21] Predictors were from the UK Biobank and included clinical data (eg, demographic, biometric and lifestyle factors).[21] The authors used a multivariable Cox model that achieved an AUC of 0.751, with machine learning methods resulting in an improvement in comparison to survival methods (random forest AUC of 0.783; deep learning AUC of 0.790).[21] The authors observed overprediction with survival models, which we also observed in the present study. The authors did not examine subgroup calibration, so it is unclear how their model performs among different population groups (eg, by income or visible minority status), which is important in understanding fairness in prediction

models. Our study was reported in line with the TRIPOD guidelines and demonstrated excellent discrimination through external validation with a c-statistic of 0.856 for the female model and 0.846 for the male model. We also examined subgroup calibration, which shows robust performance across different population characteristics. Using only self-reported survey data, PreMPoRT achieved the strongest discrimination and calibration of existing prediction models for premature mortality to date.

Implications for policy and practice

Premature mortality is an important health outcome that is frequently reported on through monitoring and surveillance initiatives to determine overall population health and to assess health system functioning.[1] The use of prediction models in applied public health settings is becoming more commonplace.[15] Prediction models may be used to identify individuals at high risk and are an important evidence-based planning tool to support targeted public health interventions.[15] This prediction model provides public health decision-makers with a tool to estimate individual risk of premature mortality using widely available population-based survey data.

Given that most premature deaths are preventable through targeted policy, intervention or treatment, this model may be used to deliver targeted or population-wide interventions to reduce premature mortality. Understanding the future distribution of risk and incidence of premature morality is important information for individuals working in both public health and healthcare. Collectively, those working in our health system play an important role in the prevention and management of chronic diseases and overall patient care, which impacts premature mortality rates. Interventions for those at high risk of premature mortality may span across the health system and could include navigation support for mental health and substance use services or health promotion strategies to encourage physical activity or smoking cessation.[52 53] Population-level public health interventions may include educational strategies to promote vaccine confidence and uptake or adaptation to the built environment to prevent pedestrian and cyclist injuries.[54 55] At present, several other prediction models developed by our team are in use across local public health departments.[35–41] Decision-makers report using these prediction models to understand the future risk of a specific health outcome and the distribution of risk within their community. The output of risk prediction models like the one described here is frequently used to support evidence-based planning and building a case for policy change or funding of public health programmes.

CONCLUSIONS

This study presents the development and validation of a novel population model for premature mortality. The model is designed to use routinely collected population-based survey data and can be used to predict the future incidence of premature mortality among populations and subgroups. This model can be used to support population-level planning and interventions, which consider the distribution of risk in the population to optimise the impact of policies and interventions. We anticipate that the model can be tested, validated and updated in populations locally and worldwide to inform strategies to reduce premature mortality.

X Laura C Rosella @LauraCRosella

Acknowledgements We would like to thank Erica Sawula from Thunder Bay District Health Unit, Julie Stratton and Kirstin Boehme from Peel Public Health and John Barbaro and Ainslie Butler from Simcoe Muskoka District Health Unit for their valuable feedback used to inform the development of PreMPoRT. This study was conducted at the Toronto RDC, a part of the Canadian Research Data Centre Network (CRDCN). This service is provided through the support of the University of Toronto, the province of Ontario, the Canadian Foundation for Innovation, the Canadian Institutes of Health Research, the Social Science and Humanity Research Council, and Statistics Canada. Additionally, Andy Hong would like to acknowledge the support from the PEAK Urban programme, funded by UKRI's Global Challenge Research Fund (Grant Ref: ES/P011055/1). All views expressed in this work are our own.

Contributors LCR is responsible for the overall content as guarantor. LCR and MO'N conceived the study. MO'N developed the methodology, performed data curation and conducted the formal analysis. MO'N also wrote the original draft of the manuscript. MH contributed to the conceptualisation of the study, performed data curation and provided input for the review and editing of the manuscript. LP contributed to the methodology and provided input for the review and editing of the manuscript. LD and KK played a role in funding acquisition, project administration and provided input for the review and editing of the manuscript. SF contributed to the methodology and provided feedback for the review and editing of the manuscript. AH contributed to funding acquisition, methodology development and provided input for the review and editing of the manuscript. DM contributed to funding acquisition, methodology development and provided input for the review and editing of the manuscript. LCR also contributed to the methodology development, supervision of the project and provided input for the review and editing of the manuscript. All authors contributed to the interpretation of the results, participated in the review process, offered valuable feedback and approved the final version of the manuscript. LCR accepted full responsibility for the finished work and/or the conduct of the study, had access to the data and controlled the decision to publish.

Funding This work was supported by the Canadian Institutes of Health Research Operating Grant (410546). LCR is supported by a Canada Research Chair (950-230702).

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Ethics approval This study was approved by the University of Toronto Research Ethics Board (Protocol #37499).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. The data used to generate the study cohort are available only through one of the Statistics Canada Research Data Centres, but restrictions apply to the availability of these data, which were used under licence for the current study, and so are not publicly available. Data are, however, available from the authors on reasonable request and with permission of the Research Data Centre.

ORCID iDs
Lief Pagalan http://orcid.org/0000-0003-2828-7085
Doug Manuel http://orcid.org/0000-0003-0912-0845
Laura C Rosella http://orcid.org/0000-0003-4867-869X

REFERENCES

1 Lewer D, Jayatunga W, Aldridge RW, *et al*. Premature mortality attributable to socioeconomic inequality in England between 2003 and 2018: an observational study. *Lancet Public Health* 2020;5:e33–41.
2 Norheim OF, Jha P, Admasu K, *et al*. Avoiding 40% of the premature deaths in each country, 2010-30: review of national mortality trends to help quantify the UN sustainable development goal for health. *Lancet* 2015;385:239–52.
3 Schroeder SA. Shattuck Lecture. We can do better--improving the health of the American people. *N Engl J Med* 2007;357:1221–8.
4 Muller DC, Murphy N, Johansson M, *et al*. Modifiable causes of premature death in middle-age in Western Europe: results from the EPIC cohort study. *BMC Med* 2016;14:87.
5 United Nations. Transforming our world: the 2030 agenda for sustainable development. 2015. Available: https://sdgs.un.org/2030agenda
6 World Health Organization. Premature mortality from noncommunicable disease. 2022. Available: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/3411
7 Canadian Institure for Health Information. Avoidable deaths. 2020. Available: https://www.cihi.ca/en/indicators/avoidable-deaths
8 Best AF, Haozous EA, Berrington de Gonzalez A, *et al*. Premature mortality projections in the USA through 2030: a modelling study. *Lancet Public Health* 2018;3:e374–84.
9 Fransham M, Dorling D. Have mortality improvements stalled in England? *BMJ* 2017;357:j1946.
10 Mackenbach JP, Bos V, Andersen O, *et al*. Widening socioeconomic inequalities in mortality in six Western European countries. *Int J Epidemiol* 2003;32:830–7.
11 Shahidi FV, Parnia A, Siddiqi A. Trends in socioeconomic inequalities in premature and avoidable mortality in Canada, 1991-2016. *CMAJ* 2020;192:E1114–28.
12 Bosworth B. Increasing Disparities in Mortality by Socioeconomic Status. *Annu Rev Public Health* 2018;39:237–51.
13 Buajitti E, Frank J, Watson T, *et al*. Changing relative and absolute socioeconomic health inequalities in Ontario, Canada: A population-based cohort study of adult premature mortality, 1992 to 2017. *PLoS One* 2020;15:e0230684.
14 Dowell SF, Blazes D, Desmond-Hellmann S. Four steps to precision public health. *Nature New Biol* 2016;540:189–91.
15 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.
16 Steyerberg EW. *Clinical Prediction Models*. 2nd edn. Springer Nature Switzerland, 2019.
17 Goldstein BA, Navar AM, Pencina MJ, *et al*. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017;24:198–208.
18 Califf RM, Harrell FE. Individual risk prediction using data beyond the medical clinic. *CMAJ* 2018;190:E947–8.
19 Manuel DG, Perez R, Sanmartin C, *et al*. Measuring Burden of Unhealthy Behaviours Using a Multivariable Predictive Approach: Life Expectancy Lost in Canada Attributable to Smoking, Alcohol, Physical Inactivity, and Diet. *PLoS Med* 2016;13:e1002082.
20 Ganna A, Ingelsson E. 5 year mortality predictors in 498,103 UK Biobank participants: a prospective population-based study. *Lancet* 2015;386:533–40.
21 Weng SF, Vaz L, Qureshi N, *et al*. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS One* 2019;14:e0214365.
22 Rosella LC, O'Neill M, Fisher S, *et al*. A study protocol for a predictive algorithm to assess population-based premature mortality risk: Premature Mortality Population Risk Tool (PreMPoRT). *Diagn Progn Res* 2020;4:18.
23 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Ann Intern Med* 2015;162:735–6.
24 Béland Y. Canadian community health survey--methodological overview. *Health Rep* 2002;13:9–14.
25 Canadian Institute for Health Information. Health indicators 2012. Ottawa, ON, 2012.
26 Remington PL, Catlin BB, Kindig DA. Monitoring Progress in Population Health: Trends in Premature Death Rates. *Prev Chronic Dis* 2013;10.
27 Thomas B, Dorling D, Smith GD. Inequalities in premature mortality in Britain: observational study from 1921 to 2007. *BMJ* 2010;341:c3639.
28 Tobias M, Jackson G. Avoidable mortality in New Zealand, 1981-97. *Aust N Z J Public Health* 2001;25:12–20.
29 Page A, Tobias M, Glover J, *et al*. Australian and New Zealand atlas of avoidable mortality. Adelaide: PHIDU, University of Adelaide; 2006.
30 Holt-Lunstad J, Smith TB, Baker M, *et al*. Loneliness and social isolation as risk factors for mortality: a meta-analytic review. *Perspect Psychol Sci* 2015;10:227–37.
31 Ouellet BL, Romeder J-M, Lance J-M. Premature mortality attributable to smoking and hazardous drinking in Canada. *Am J Epidemiol* 1979;109:451–63.
32 Stringhini S, Carmeli C, Jokela M, *et al*. Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1·7 million men and women. *Lancet* 2017;389:1229–37.
33 Khan AM, Urquia M, Kornas K, *et al*. Socioeconomic gradients in all-cause, premature and avoidable mortality among immigrants and long-term residents using linked death records in Ontario, Canada. *J Epidemiol Community Health* 2017;71:625–32.
34 Buajitti E, Watson T, Norwood T, *et al*. Regional variation of premature mortality in Ontario, Canada: a spatial analysis. *Popul Health Metr* 2019;17:9.
35 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.
36 Taljaard M, Tuna M, Bennett C, *et al*. Cardiovascular Disease Population Risk Tool (CVDPoRT): predictive algorithm for assessing CVD risk in the community setting. A study protocol. *BMJ Open* 2014;4:e006701.
37 Fisher S, Hsu A, Mojaverian N, *et al*. Dementia Population Risk Tool (DemPoRT): study protocol for a predictive algorithm assessing dementia risk in the community. *BMJ Open* 2017;7:e018018.
38 Manuel DG, Tuna M, Perez R, *et al*. Predicting Stroke Risk Based on Health Behaviours: Development of the Stroke Population Risk Tool (SPoRT). *PLoS One* 2015;10:e0143342.
39 Rosella LC, Kornas K, Yao Z, *et al*. Predicting High Health Care Resource Utilization in a Single-payer Public Health Care System: Development and Validation of the High Resource User Population Risk Tool. *Med Care* 2018;56:e61–9.
40 Lebenbaum M, Espin-Garcia O, Li Y, *et al*. Development and validation of a population based risk algorithm for obesity: The Obesity Population Risk Tool (OPoRT). *PLoS One* 2018;13:e0191169.
41 Ng R, Sutradhar R, Kornas K, *et al*. Development and Validation of the Chronic Disease Population Risk Tool (CDPoRT) to Predict Incidence of Adult Chronic Disease. *JAMA Netw Open* 2020;3:e204669.
42 van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* 2007;16:219–42.
43 Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 1987.
44 Oksuzyan A, Juel K, Vaupel JW, *et al*. Men: good health and high mortality. Sex differences in health and aging. *Aging Clin Exp Res* 2008;20:91–102.
45 Rosella LC, Calzavara A, Frank JW, *et al*. Narrowing mortality gap between men and women over two decades: a registry-based study in Ontario, Canada. *BMJ Open* 2016;6:e012564.
46 Liang CY, Buajitti E, Kornas K, *et al*. Canadian atlas of premature mortality, 1992-2015: cross-provincial comparisons. 2020.
47 Harrell FE. *Regression Modeling Strategies*. Cham: Springer International Publishing, 2015:3.

48 Kovacevic MS, Mach L, Roberts G. Bootstrap variance estimation for predicted individual and population-average risks. 2008. Available: https://www.semanticscholar.org/paper/Bootstrap-Variance-Estimation-for-Predicted-and-Kova%C4%8Devi%C4%87-Mach/cbe8e6d3fb8b548d0dd925acd1eef3a974a28a83?p2df
49 Fisher S, Bennett C, Hennessy D, *et al*. International population-based health surveys linked to outcome data: A new resource for public health and epidemiology. *Health Rep* 2020;31:12–23.
50 Park J, Tjepkema M, Goedhuis N, *et al*. Avoidable mortality among First Nations adults in Canada: A cohort analysis. *Health Rep* 2015;26:10–6.
51 Kouyoumdjian FG, Kiefer L, Wobeser W, *et al*. Mortality over 12 years of follow-up in people admitted to provincial custody in Ontario: a retrospective cohort study. *CMAJ Open* 2016;4:E153–61.
52 Waid J, Halpin K, Donaldson R. Mental health service navigation: a scoping review of programmatic features and research evidence. *Soc Work Ment Health* 2021;19:60–79.
53 Ussher MH, Faulkner GEJ, Angus K, *et al*. Exercise interventions for smoking cessation. *Cochrane Database Syst Rev* 2019;2019:CD002295.
54 Rothman L, Schwartz N, Cloutier M-S, *et al*. Child pedestrian and cyclist injuries, and the built and social environment across Canadian cities: the Child Active Transportation Safety and the Environment Study (CHASE). *Inj Prev* 2022;28:311–7.
55 Rashid H, Yin JK, Ward K, *et al*. Assessing Interventions To Improve Influenza Vaccine Uptake Among Health Care Workers. *Health Aff* 2016;35:284–92.
56 Visible minority of person [Statistics Canada]. Government of Canada. Available: https://www23.statcan.gc.ca/imdb/p3Var.pl?Function=DEC&Id=45152 [Accessed 25 Aug 2024].