Diagnostic and Prognostic Research

PROTOCOL

Open Access

A study protocol for a predictive model to assess population-based avoidable hospitalization risk: Avoidable Hospitalization Population Risk Prediction Tool (AvHPoRT)

Laura C. Rosella[1,2,3,4*], Mackenzie Hurst[1,4], Meghan O'Neill[1], Lief Pagalan[1], Lori Diemert[1], Kathy Kornas[1], Andy Hong[5,6,7], Stacey Fisher[1,8] and Douglas G. Manuel[8,9,10,11,12]

Abstract

Introduction Avoidable hospitalizations are considered preventable given effective and timely primary care management and are an important indicator of health system performance. The ability to predict avoidable hospitalizations at the population level represents a significant advantage for health system decision-makers that could facilitate proactive intervention for ambulatory care-sensitive conditions (ACSCs). The aim of this study is to develop and validate the Avoidable Hospitalization Population Risk Tool (AvHPoRT) that will predict the 5-year risk of first avoidable hospitalization for seven ACSCs using self-reported, routinely collected population health survey data.

Methods and analysis The derivation cohort will consist of respondents to the first 3 cycles (2000/01, 2003/04, 2005/06) of the Canadian Community Health Survey (CCHS) who are 18–74 years of age at survey administration and a hold-out data set will be used for external validation. Outcome information on avoidable hospitalizations for 5 years following the CCHS interview will be assessed through data linkage to the Discharge Abstract Database (1999/2000–2017/2018) for an estimated sample size of 394,600. Candidate predictor variables will include demographic characteristics, socioeconomic status, self-perceived health measures, health behaviors, chronic conditions, and area-based measures. Sex-specific algorithms will be developed using Weibull accelerated failure time survival models. The model will be validated both using split set cross-validation and external temporal validation split using cycles 2000–2006 compared to 2007–2012. We will assess measures of overall predictive performance (Nagelkerke R^2), calibration (calibration plots), and discrimination (Harrell's concordance statistic). Development of the model will be informed by the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement.

Ethics and dissemination This study was approved by the University of Toronto Research Ethics Board. The predictive algorithm and findings from this work will be disseminated at scientific meetings and in peer-reviewed publications.

Keywords Prediction model, Avoidable hospitalization, Ambulatory care sensitive conditions, Study protocol, Population level, Survival analysis

*Correspondence:
Laura C. Rosella
laura.rosella@utoronto.ca
Full list of author information is available at the end of the article

Strengths and limitations

- The Avoidable Hospitalization Population Risk Tool (AvHPoRT) will use routinely collected population-based survey data individually linked to health administrative data in Canada to develop and validate a risk prediction tool for avoidable hospitalizations associated with ambulatory care sensitive conditions
- AvHPoRT will improve existing risk prediction tools for avoidable hospitalization by encompassing non-medical determinants of health such as self-reported demographic characteristics, socioeconomic status, health behaviors, and area-based measures
- Because this model includes non-medical data, we can predict at the population level social determinants of health factors before individuals enter the hospital system, making it useful for public health-focused applications. This addition is a distinct advantage over existing hospital-based algorithms primarily used for triaging people that are already in contact with the acute care system
- The proposed analytic plan follows the recommendations published in the TRIPOD statement for multivariable predictive models to reduce statistical overfitting
- Despite a robust validation approach, including both split set validation and external temporal validation, further validation may be necessary to assess generalizability and calibration for applications outside of Canada
- AvHPoRT can be leveraged by health system decision-makers and planners to identify subgroups of the population at high risk of avoidable hospitalization, to inform population management and prevention approaches, and to estimate the future burden of avoidable hospitalizations in Canada

Background

Avoidable hospitalizations refer to hospitalizations for conditions that can be prevented, treated, or managed in primary care and, therefore should not necessitate hospitalization [1, 2]. This set of conditions is typically referred to as ambulatory care-sensitive conditions (ACSCs). In the Canadian context, avoidable hospitalizations are defined to include any acute care hospitalization among individuals 0–74 years of age for any of seven ACSCs, including angina, asthma, congestive heart failure, chronic obstructive pulmonary disease, diabetes, epilepsy, and hypertension where the patient is alive at discharge [3]. Variations of this definition exist across health systems, including acute cellulitis, dental conditions, vaccine-preventable conditions (e.g., influenza), and alternative age specifications [4]. Avoidable hospitalizations are an important health system performance indicator that signals poor management of health conditions [5, 6] and inadequate access to quality preventive care in the community [3]. The Canadian Institutes for Health Information (CIHI) estimated that 6.8 million Canadians aged 20–74, have an avoidable hospitalization resulting in approximately 95,000 hospitalizations and 13,000 deaths per year [7, 8]. Despite Canada's universal healthcare system covering all medically necessary services, social, sex, and geographic inequalities in avoidable hospitalizations exist [7, 9–11]. Additionally, avoidable hospitalizations are expensive for the healthcare system and in 2006, avoidable hospitalizations were estimated to cost the Canadian healthcare system $416 million annually [11].

Several risk factors have been associated with hospitalization for ACSCs, including demographics [12–16], health behaviors [12, 16–18], rurality of residence [6, 13, 19–23], socioeconomic status [9, 10, 16, 24–27], chronic conditions [12, 28, 29], and characteristics related to the organization, structure, and delivery of care [6, 30–33]. Clinical models have been developed to predict the risk of emergency and inpatient hospitalization or re-hospitalization; however, none have specifically been developed for avoidable hospitalizations at the population level and for the Canadian context [34–37]. Canadian researchers have previously developed a simple and complex version of the Hospital Admission Risk Prediction tool to identify those at risk of future 30-day and 15-month all-cause hospitalization using administrative data from Ontario and Manitoba [38] based on hospital utilization variables, achieving moderate discrimination (*c*-statistic 0.66–0.70). Given that survey data were not used, details on socio-demographics (e.g., immigration status, ethnicity), individual-level socioeconomic status (e.g., income, education), social support (e.g., living alone), or health behaviors (e.g., smoking, alcohol consumption, body weight) were not included. Therefore, there are opportunities to improve model performance by adding variables that are lacking in administrative data and creating a model that better informs population health approaches [39]. Furthermore, a model that allows for prediction at the population level contributes by allowing for accurate distribution of risk in the population, which can ensure that strategies for prevention are allocated to the populations that will most likely benefit from attention and outreach. This approach allows resources to be directed in a way that addresses risk along multiple determinants of health and optimizes the impact on populations [40]. This is a critical cornerstone of population health management, an approach increasingly being adopted by many health systems [41].

To address the need for a population-based prediction model for avoidable hospitalizations that encompasses non-medical determinants of health, we propose the development and validation of the Avoidable Hospitalization Population Risk Tool (AvHPoRT). AvHPoRT will use self-reported risk factor information for population-based risk prediction of the first avoidable hospitalization in adults in Canada for seven ACSCs over 5 years: angina, asthma, congestive heart failure, COPD, diabetes, epilepsy, and hypertension. According to recommendations in TRIPOD guidelines, this study protocol pre-specifies the predictor variables and analytic plan for the development and validation of AvHPoRT [42].

Methods and analysis

Data sources

Canadian Community Health Survey

The Canadian Community Health Survey (CCHS) is a cross-sectional survey that collects information on self-reported sociodemographic characteristics, personal health status, health behaviors, and healthcare utilization for the Canadian population aged 12 years and older [43]. The study base is all Canadian youth and adults, excluding individuals living in certain remote regions of Quebec and Nunavut, full-time members of the Canadian Forces, persons living on reserves and other Aboriginal settlements, and individuals living in institutions [43]. Overall, these exclusions represent < 3% of the Canadian population [43]. Canadian respondents from the following CCHS cycles–1.1 (2000/01), 2.1 (2003/04), 3.1 (2005/06), 2007/08, 2009/10, and 2011/12 will be used to create the overall study cohort and obtain predictor variables. The estimated sample size 394,600.

Discharge abstract database and Canadian Vital Statistics Database

The Discharge Abstract Database (DAD) is a national database maintained by the Canadian Institutes of Health Research (CIHI) that contains information on demographic, administrative, and clinical data on hospital inpatient discharges and same-day surgery procedures [44]. All provinces and territories, excluding Quebec, submit information to DAD, representing 75% of all inpatient hospital discharges in Canada. We will use the most recent CCHS-DAD linkage, including data from the fiscal year (FY) 1999/2000–2017/2018. DAD will be used to identify all hospital-based deaths and avoidable hospitalization for seven ACSCs. The Canadian Vital Statistics Database (CVSD) is a national database that includes all deaths registered in Canada, including information on the death date and cause of death. In hospitals, deaths are captured in the DAD. CCHS respondents consented to linkage to health administrative databases, resulting in a linkage rate between the CCHS-DAD and CCHS-CVSD of approximately 85% (excluding Quebec respondents who do not participate) [45, 46]. Further details on the linkage process are described elsewhere [45].

Canadian Marginalization Index

The Canadian Marginalization Index (CAN-Marg) is a census-based measure of sociodemographic characteristics, including households and dwellings, material resources, age and labor force, and immigration and visible minority status at the dissemination area level [47]. The dissemination area is the smallest geographic unit for which census information is available, representing approximately 200–700 persons [48]. The 2001, 2006, and 2011 CAN-Marg indices will be used [49]. CAN-Marg is linked to respondents in the CCHS based on their postal code at the time of survey administration.

Patient and public involvement

We have consulted with partners in public health units across urban and rural Ontario in the development of several population-based risk tools. Our partners at local public health units in Ontario provided feedback on the present protocol, informing the relevant candidate variables and how they would use AvHPoRT once validated. Individual patients will not be involved in the design, conduct, reporting, or dissemination of this research.

Study design

Using population-based survey data linked to health administrative data, sex-specific AvHPoRT models will be developed and validated using CCHS respondents from the survey years 2000–2012. All analyses will be sex-stratified due to differences in the individual risk factors for first avoidable hospitalizations found in previous studies on this population [16]. As a result, the models will be sex-specific. Two development and validation approaches will be used: split set validation and external validation on a hold-out dataset. Individuals will be followed for 5 years after the CCHS interview date until the first avoidable hospitalization, death, or end of the study period, whichever comes first. For both the development and validation cohorts, respondents will be excluded if they (1) are less than 18 years of age or older than 74 years of age, (2) live in Quebec or the Territories or, (3) are pregnant at the time of the CCHS interview. The lower bound age limitation is due to the differing nature of youth healthcare utilization who are typically under parental or legal guardian care. The upper age limit is due to how CIHI defines avoidable hospitalizations for ACSCs [3]. The rationale is that hospitalizations for the seven conditions captured by definition are deemed less avoidable or completely unavoidable after age 74

due to declines in overall health [3]. Quebec residents will be excluded because this province does not submit DAD records and the Territories will be excluded due to high levels of missing data [45]. Additionally, pregnant respondents will be excluded due to the inability to correctly estimate body mass index (BMI) and the potential for misclassification of baseline covariates (i.e., smoking or alcohol consumption status).

Identification of potential predictor variables

Predictor variables from the CCHS capture baseline information of the study cohort. We selected variables based on their availability across provinces and cycles, reviewed existing literature on avoidable hospitalization risk factors, observational studies based on survey data linked to avoidable hospitalizations in Canada [16, 27], recommendations from knowledge users in public health, and expertise from our team with prior experience developing and validating predictive algorithms for healthcare use [50], acute and chronic conditions [51–56] and mortality [57, 58]. We limited to variables consistent across provinces and CCHS survey cycles. After screening available predictors, a total of 39 candidate variables were selected, including demographic characteristics (e.g., age, ethnicity), socioeconomic variables (e.g., immigration status, marital status, household income, education), self-perceived measures (e.g., general health, life stress, and community belonging), health behaviors (e.g., cigarette smoking, alcohol consumption, fruit and vegetable consumption, physical activity, BMI), chronic conditions (e.g., diabetes, COPD), healthcare access (e.g., whether respondent has a family physician), use preventive care (ever had a flu shot), and five area-based measures (e.g., four CAN-Marg Indices and one CCHS based individual measure).

Outcome

Avoidable hospitalization was defined as a hospitalization among adults between 18 and 74 years of age at the time of admission, where discharged alive, and the most responsible diagnosis was one of seven chronic ACSCs: angina (excluding certain cardiac interventions), asthma, congestive heart failure (excluding certain cardiac interventions), chronic obstructive pulmonary disease, diabetes, epilepsy, and hypertension (excluding certain cardiac interventions) (Table 1) [3]. We will consider the first avoidable hospitalization as the outcome. CIHI defines the most responsible diagnosis as the most responsible condition for the patient's stay in the facility [59]. Per the CIHI definition, certain cardiac interventions performed for hospitalizations due to angina, congestive heart failure, and hypertension are identified using the International Classification of Diseases-9 and the International Classification of Diseases-10 diagnostic codes and will be excluded [3, 60].

Sample size

To increase the likelihood of developing a robust prediction model, we will combine multiple cycles of the CCHS to increase the sample size. In taking this approach, we aim to minimize the potential for model overfitting and to increase the precision of predictions. A recent Canadian study by our team examined individual and area-level factors that were associated with the risk of avoidable hospitalizations, which included the first 6 cycles of the CCHS linked to the DAD with study exclusions that are consistent with those proposed in this protocol [27]. Therefore, we anticipate approximately 389,100 respondents with 8,500 (2.2%) avoidable hospitalizations [16]. Following CCHS sampling methods [60], we anticipate an approximately equal number of males and females, resulting in approximately 195,000 observations for each sex-specific model.

We calculated the minimum sample size for AvH-PoRT using the *pmsampsize* package in R according to the approach proposed by Riley et al., which considers context-specific factors including the total number of study participants, the proportion of the outcome in the study population, and the anticipated predictive performance of the model [61, 62]. Assuming a Nagelkerke's R of 16.5% based on a comparable risk prediction tool for Ontario [38], with an outcome proportion of 2.1%, 163 parameters (82 from all variables in Table 2 plus 81 for testing interactions with linear age), and a shrinkage factor of 0.90 the minimum sample size is 16,664 respondents with 333 events per model. The expected sample size surpasses these minimum estimates, and as a result, we anticipate a sufficient sample for our proposed analyses.

Analysis plan

The analytic plan was developed following the guidelines for prediction models by Steyerberg [63] and Harrell [64]. This plan was constructed after accessing our study cohort but before model fitting or evaluation of descriptive statistics examining relationships between predictor variables and the outcome. Important considerations that informed the analytic approach include full pre-specification of predictor variables and implementation of flexible functions for continuous variables. As a final step, we will verify the sequential addition of predictors using the least absolute shrinkage and selection operator (LASSO) [65]. In the case of categorical variables, if the LASSO selects only one of the values in a categorical variable, we will manually test after to see if keeping the variable in the model or removing it will give better model performance. In

Table 1 ICD-9 and ICD-10 codes used to capture avoidable hospitalizations as defined by the Canadian Institutes for Health Information

Condition	ICD 9/9-CM and ICD-10-CA
Grand mal status and other epileptic convulsions	ICD-9/9-CM: 345 ICD-10-CA: G40, G41
Chronic obstructive pulmonary disease	Any most responsible diagnosis code of ICD-9/9-CM: 491, 492, 494, 496 ICD-10-CA: J41, J42, J43, J44, J47 Most responsible diagnosis of acute lower respiratory infection, only when a secondary diagnosis of J44 in ICD-10-CA or 496 in ICD-9/9-CM is also present ICD-9/9-CM: 466, 480–486, 487.0 ICD-10-CA: J10.0, J11.0, J12–J16, J18, J20, J21, J22
Asthma	ICD-9/9-CM: 493 ICD-10-CA: J45
Diabetes	ICD-9: 250.0, 250.1, 250.2, 250.7 ICD-9-CM: 250.0, 250.1, 250.2, 250.8 ICD-10-CA: E10.0, E10.1, E10.63, E10.64, E10.9 E11.0, E11.1, E11.63, E11.64, E11.9 E13.0, E13.1, E13.63, E13.64, E13.9 E14.0, E14.1, E14.63, E14.64, E14.9
Heart failure and pulmonary edema	ICD-9/9-CM: 428, 518.4 ICD-10-CA: I50, J81
Hypertension	ICD-9/9-CM: 401.0, 401.9, 402.0, 402.1, 402.9 ICD-10-CA: I10.0, I10.1, I11
Angina	ICD-9: 411, 413 ICD-9-CM: 411.1, 411.8, 413 ICD-10-CA: I20, I23.82, I24.0, I24.8, I24.9
Cardiac procedure codes for exclusion	Codes for exclusion applied to heart failure and pulmonary edema, hypertension, and angina: CCP: 47^^, 480^–483^, 489.1, 489.9, 492^–495^, 497^, 498^ ICD-9-CM: 336, 35^^, 36^^, 373^, 375^, 377^, 378^, 379.4–379.8 CCI: 1.HA.58.^^, 1.HA.80.^^, 1.HA.87.^^, 1.HB.53.^^, 1.HB.54.^^, 1.HB.55.^^, 1.HB.87.^^, 1.HD.53.^^, 1.HD.54.^^, 1.HD.55.^^, 1.HH.59.^^, 1.HH.71.^^, 1.HJ.76.^^, 1.HJ.82.^^, 1.HM.57.^^, 1.HM.78.^^, 1.HM.80.^^, 1.HN.71.^^, 1.HN.80.^^, 1.HN.87.^^, 1.HP.76.^^, 1.HP.78.^^, 1.HP.80.^^, 1.HP.82.^^, 1.HP.83.^^, 1.HP.87.^^, 1.HR.71.^^, 1.HR.80.^^, 1.HR.84.^^, 1.HR.87.^^, 1.HS.80.^^, 1.HS.90.^^, 1.HT.80.^^, 1.HT.89.^^, 1.HT.90.^^, 1.HU.80.^^, 1.HU.90.^^, 1.HV.80.^^, 1.HV.90.^^, 1.HW.78.^^, 1.HW.79.^^, 1.HX.71.^^, 1.HX.78.^^, 1.HX.79.^^, 1.HX.80.^^, 1.HX.83.^^, 1.HX.86.^^, 1.HX.87.^^, 1.HY.85.^^, 1.HZ.53 rubric (except 1.HZ.53.LA-KP), 1.HZ.54.^^, 1.HZ.55 rubric (except 1.HZ.55.LA-KP), 1.HZ.56.^^, 1.HZ.57.^^, 1.HZ.59.^^, 1.HZ.80.^^, 1.HZ.85.^^, 1.HZ.87.^^, 1.IF.83.^^, 1.IJ.50.^^, 1.IJ.54.GQ-AZ, 1.IJ.55.^^, 1.IJ.57.^^, 1.IJ.76.^^, 1.IJ.80.^^, 1.IJ.86.^^, 1.IK.50.^^, 1.IK.57.^^, 1.IK.80.^^, 1.IK.87.^^, 1.IN.84.^^, 1.LA.84.^^, 1.LC.84.^^, 1.LD.84.^^, 1.YY.54.LA-NJ, 1.YY.54.LA-FS, 1.YY.54.LA-NM, 1.YY.54.LA-FR, 1.YY.54.LA-FU

^: Any single additional character

^^: Any two additional characters

agreement with recent publications that have called for improvements to the design and reporting of prediction models [66], we have proposed this study protocol to help improve the transparency of the model-building process, increase the robustness of our prediction model, and limit type I errors. Data cleaning and coding of predictors will be conducted in SAS V.9.4, and model development and validation will be carried out in R using Harrell's *HMisc* [67] and *rms* packages of functions, *'survey'* among others [68]. This protocol was developed following recommendations of the TRIPOD statement for multivariable predictive models, which will also inform the reporting of AvHPoRT [42, 69].

Coding and cleaning of predictor variables

All data cleaning and coding of predictor variables will occur prior to examining exposure-outcome relationships. Descriptive statistics and boxplots will be used to examine the width of distributions. Continuous variables with highly skewed distributions will be winsorized to the 99.5th percentile, which will set all extreme values or outliers to the 99.5th percentile. We will also take into account how predictor variables have been modeled in prior risk prediction models [50–55, 57]. Consistent with prior model development, some predictor variables will be derived based on a combination of variables in the

Table 2 Pre-specification of predictor variables for AvHPoRT with initial degrees of freedom (df)

Variable grouping	Variable	Definition	df	Scale
Demographics	**Age** (years)	5-knot restricted cubic spline	4	Continuous
	Self-identified ethnicity		1	Dichotomous
	White			
	Non-white			
Socioeconomic status and self-perceived measures	**Immigration status**		2	Categorical
	Canadian born			
	Recent immigrant	Immigrated < 10 years		
	Non-recent immigrant	Immigrated ≥ 10 years		
	Marital status		2	Categorical
	Single never married			
	Domestic partner (married/common law)			
	Widowed/separated/divorced			
	Household income		4	Ordinal
	Quintile 1	Lowest 20%		
	Quintile 2			
	Quintile 3			
	Quintile 4			
	Quintile 5	Highest 20%		
	Education		3	Ordinal
	Less than secondary school graduation			
	Secondary school graduation			
	Some post-secondary education			
	Post-secondary completed			
	Self-perceived general health		4	Ordinal
	Excellent			
	Very good			
	Good			
	Fair			
	Poor			
	Self-perceived life stress		4	Ordinal
	Not at all			
	Not very			
	A bit stressful			
	Quite a bit			
	Extremely stressful			
	Self-perceived community belonging		3	Ordinal
	Very strong			
	Somewhat strong			
	Somewhat weak			
	Very weak			

Table 2 (continued)

Variable grouping	Variable	Definition	df	Scale
Health behaviors	**Cigarette smoking**		4	Categorical
	Non-smoker	Never a smoker or former occasional smoker with < 100-lifetime cigarettes		
	Former heavy smoker	Former smoker [≥ 1 pack (25 cigarettes)/day]		
	Former light smoker	Former smoker [< 1 pack (25 cigarettes)/day]		
	Heavy smoker	Current smoker [≥ 1 pack (25 cigarettes)/day]		
	Light smoker	Current smoker [< 1 pack (25 cigarettes)/day]		
	Alcohol consumption		3	Categorical
	Non-drinker	No alcohol consumption in the last 12 months or drink frequency fewer than once a week		
	Light drinker	Alcohol consumption frequency at least once a week and 0–2 (females) or 0–3 (males) drinks in the previous week		
	Moderate drinker	3–14 (females) or 4–21 (males) drinks in the previous week		
	Heavy drinker	≥ 14 (females) or ≥ 21 (males) drinks in the previous week, or binging behavior on a weekly basis (≥ 5 drinks on any occasion)		
	Daily fruit and vegetable consumption		2	Ordinal
	Low consumption	0 to less than 3 times daily		
	Medium consumption	3 to less than 6 times daily		
	High consumption	6 or more times daily		
	Leisure physical activity (kcal/kg/day)		2	Ordinal
	Active	3.0 or more metabolic equivalents per day		
	Moderate	1.5–2.9 metabolic equivalents per day		
	Inactive	Less than 1.5 metabolic equivalents per day		
	Body mass index (BMI) (kg/m^2)	5-knot restricted cubic spline	4	Continuous
Chronic conditions	**Self-reported chronic conditions diagnosed by a health professional**		17	Dichotomous
	Including asthma, arthritis, back problems, high blood pressure, migraines, emphysema, chronic obstructive pulmonary disease, diabetes, heart disease, cancer, intestinal ulcers, stroke, urinary incontinence, bowel disease, mood disorder, or anxiety disorder	Yes/no for each individual chronic condition		
Healthcare utilization	Ever had a flu shot	Yes/no	1	Binary
	Has a regular medical doctor	Yes/no	1	Binary

Table 2 (continued)

Variable grouping	Variable	Definition	df	Scale
Area-based variables	**Rurality**		1	Dichotomous
	Population center	Population of at least 1000 and a density of ≥ 400 people per square kilometer based on current census population counts		
	Rural area	Population concentration or densities below the urban threshold based on current census population counts		
	Material deprivation		4	Ordinal
	Quintile 1	Least deprived		
	Quintile 2			
	Quintile 3			
	Quintile 4			
	Quintile 5	Most deprived		
	Ethnic diversity		4	Ordinal
	Quintile 1	Least diverse		
	Quintile 2			
	Quintile 3			
	Quintile 4			
	Quintile 5	Most diverse		
	Residential instability		4	Ordinal
	Quintile 1	Least unstable		
	Quintile 2			
	Quintile 3			
	Quintile 4			
	Quintile 5	Most unstable		
	Dependency		4	Ordinal
	Quintile 1	Least dependent		
	Quintile 2			
	Quintile 3			
	Quintile 4			
	Quintile 5	Most dependent		
	Area-level household income		4	Ordinal
	Quintile 1	Lowest		
	Quintile 2			
	Quintile 3			
	Quintile 4			
	Quintile 5	Highest		

CCHS. For example, alcohol consumption will be defined based on a combination of three variables, including whether a respondent reported drinking in the past year, the number of times the respondent drank in the past week, and the total number of drinks consumed in the past week to create a final variable with four categories.

A BMI correction equation will be used to reduce the bias in self-reported height and weight [70]. All predictor variables and their definitions have been pre-specified to minimize the possibility of overfitting (Table 2). Additional details on the CCHS questions used to create health risk behavior variables are available as a supplementary file (see online Supplementary file 1).

Approach to missing data

Given the limitations associated with complete case analysis, including inefficiency and selection bias, multiple imputations will be used to assign values to missing predictor variables using the *mice* package in R, which imputes incomplete multivariate data by chained equations (mice) [71, 72]. Using the mice procedures [73], we will incorporate the full list of predictors, the outcome, and auxiliary variables (i.e., variables that are not predictors but may be useful in lending information to impute missing values) in the imputation procedure. A total of up to five imputed datasets will be generated. The final model will be run on each imputed dataset separately, and the results will be combined using the rules recommended by Rubin and Schenker [72] to account for imputation uncertainty. Missing rates in the data source are known to be rather low (< 5%) and hence no attempts are made to check sensitivity to the MAR assumption. The assessment of model performance based on multiple imputations is a challenging task and we will closely follow the guidance provided by Wood et al. [74].

Model specification

Sex-specific models will be developed using the pre-specified predictor variables outlined in Table 2. Continuous predictors will be modeled flexibly using restricted cubic splines with piecewise cubic functions smoothed at the knot placements based on Harrell's percentile recommendations [64]. Table 2 presents the initial model specification which includes 82 degrees of freedom. During the model-building process, we will also examine alternate variable forms used in prior models to perform best. For example, we aim to include BMI as both a continuous predictor (i.e., body mass index as specified in Table 2); however, will also test BMI in its ordinal form using the World Health Organization classifications for BMI (underweight, normal weight, overweight, obese type I, obese type II, and obese type III). We will compare the continuous form with the categorical form using measures of overall predictive performance, model fit, discrimination, calibration, and information criterion (e.g., AIC and BIC). The form of the predictor that improves the overall model fit will be chosen for the final model. Variables will be centered on their means for ease of recalibration in new populations, which can center data on local means. In addition, we will consider interaction terms with linear age and all other variables listed in Table 2.

Initially, we will fit a full multivariable model containing all prior predictors as specified in Table 2. Then as a subsequent step, we will apply the least absolute shrinkage and selection operator (LASSO) will be used for variable selection [75]. Since the value of lambda plays a very important role for LASSO, a k-fold cross-validation method will be used on the derivation cohort to select the appropriate value of lambda by comparing the partial-likelihood deviance [74]. Should predictors be selected in one imputed dataset and not the other, careful consideration will be made to decide which predictors will be chosen for the final model. For example, to build a more comprehensive model, both predictors may be kept; however, if both predictors are from the same casual path, then we may choose only one so they do not interfere with each other and reduce the model performance. Each predictor will be carefully considered in this way and documented in the final paper for AvHPoRT.

Model estimation

The 5-year risk of having an avoidable hospitalization will be assessed using sex-specific Weibull accelerated failure time survival models. Our team's previous work using development and validation methods has demonstrated that Weibull models perform well for population-based prediction models [49–54]. Using a survival model will also properly handle premature mortality, which is the competing risk of avoidable hospitalizations, by censoring individuals once they have had a premature mortality.

To confirm the parametric assumptions of the Weibull model are met, stratified Kaplan-Meier estimates will be made and log-log plots will be plotted against log survival time to confirm they are approximately linear and parallel [62]. The proportional hazards assumption will be checked by plotting stratified log cumulative hazards and assessing the Schoenfeld residuals. Predictor-time interaction terms will be added to the model if required. To ensure the proposed analysis is representative of the Canadian population, survey weights provided by Statistics Canada will be used to account for complexities in the CCHS survey design.

Model validation

The model will first be derived using the first three CCHS cycles ((1.1 (2000/01), 2.1 (2003/04), 3.1 (2005/06)). We will internally validate using a split sample approach where the 70% development model will be applied to the remaining 30%. The model will then be externally validated in a hold-out dataset using all data of the last three CCHS cycles (2007/08, 2009/10, and 2011/12). Once the

final model is determined, all data will be combined to estimate the final application of the model. In order to assess if the model performs similarly across age, income quintile, region, sex, immigration, and education, we will examine performance across geography, levels of education, income, and immigration status.

Assessment of model performance

The overall predictive performance in the derivation, validation, and combined cohorts will be examined and reported according to overall measures of predictive accuracy, discrimination, and calibration. Measures of overall predictive accuracy include the proportion of variance explained by predictive variables (i.e., Nagelkerke's R^2) and the Integrated Brier score [76]. Discrimination is defined as the ability of a model to correctly differentiate between respondents who develop the outcome vs respondents who do not [76]. Discrimination will be evaluated using Harrell's concordance statistic with 95% confidence intervals estimated using bootstrap samples. In the evaluation of predictive performance, Steyerberg [63] and Cook [77, 78] recommend routine assessment of calibration and calibration slopes. Model calibration will primarily be evaluated by visually comparing the observed and predicted risk of avoidable hospitalization over deciles of predicted risk using calibration points over different periods of time (e.g., 1, 3, and 5 years). We will prioritize visual inspection of calibration plots which is less influenced by a large sample size, in contrast to formal statistical significance testing [79]. Calibration slopes will be created by regressing the outcome in the validation cohort on the predicted risk of avoidable hospitalization, thereby reflecting true differences in the effects of predictor variables and the effect of overfitting to the development cohort. Perfect calibration is indicated by a slope of 1 which will be evaluated using the Wald or likelihood ratio tests. Adequate calibration across subgroups defined above will be defined as a relative difference of less than 20% between observed and predicted risk within subgroups. The distribution of the risk of avoidable hospitalization will be assessed for extreme values and outliers and clearly reported in our final paper.

Model presentation

The final AvHPoRT model will be presented with both beta coefficients as well as hazard ratios and corresponding 95% confidence intervals. With population-based predictions as the primary goal, the model presentation will also include the model coefficients. In addition, a figure of 5-year risk across all individuals will be generated to describe the distribution of risk of avoidable hospitalization. The planning and dissemination of our model is informed by the Population Health Planning Knowledge-to-Action Model [80] developed and evaluated by our team [81]. Once the model is validated, we will carry out training workshops to build health system capacity in a local context where the model is being used. Specifically, we will develop training workshops that our team holds with local public health units to demonstrate how risk prediction tools can be leveraged to inform decision-making and planning in their setting. To increase the accessibility of AvHPoRT, the program to run the model will be made available in several statistical packages and formats, including user-friendly, point-and-click web applications. It is important to note that the use of this model to plan interventions should be accompanied by a careful evaluation of the benefit achieved, preferably accompanied by high-quality evidence on the efficacy of the proposed interventions. Furthermore, the variables that are used in the risk assessment may also be known to healthcare professionals taking care of patients or the general population. Individuals, health care providers, and system planners may act on the information from the variables in the models in different ways all contributing to outcomes in the population.

Discussion

Avoidable hospitalizations are an important health system indicator that is meaningful in the context of health system evaluation and improvement. Existing risk prediction tools for avoidable hospitalizations and other similar endpoints (i.e., emergency, inpatient, and re-admission) have not been developed using population survey data and do not contain important modifiable risk factors such as socioeconomic status and health behaviors. Importantly, existing tools rely on data from individuals once they have already entered the acute care system, which does not support public health prevention approaches that are often led by public health outside the acute care sector. Currently, health decision-makers across the health system do not have a simple and streamlined approach to estimate the incidence of avoidable hospitalization tailored to their local populations, which can further complement efforts at the individual level. The ability to accurately predict the incidence of avoidable hospitalizations at the population level using modifiable risk factors will inform both broad and targeted prevention approaches and support population health management. The purpose of our model is to estimate the risk of the first avoidable hospitalization within a 5-year period as that is the indicator defined in the Canadian context where the model will be used. Future models can also consider subsequent avoidable hospitalizations using

survival models that consider multiple events. This is a future application not covered in this analysis.

Limitations

There are some limitations to the proposed development and validation of AvHPoRT. First, the study population will be limited to CCHS respondents who agreed to share and link their responses (>80% of respondents) with the DAD. To accommodate for these small underlying differences between the subset of respondents who agreed to share their responses and the full CCHS cohort, we will use survey weights provided by Statistics Canada [82]. Additionally, the data that will be used to develop AvHPoRT is based on self-reported predictors captured at a single point in time with the potential for systematic and non-directional misclassification error. Despite this, variables from self-reported CCHS data have produced robust prediction models in the past [50–55, 57]. Furthermore, the main reason to use such data is that it is regularly and widely available to health planners. While we anticipate that AvHPoRT will be representative of the majority of the Canadian adult population (98%), some groups are not captured in the CCHS sampling frame, including on-reserve Indigenous peoples. This is an important consideration because persons of Indigenous identity have been reported to have higher rates of avoidable hospitalizations in Canada [83]. Due to limits in data sharing and availability, residents of Quebec and the Territories will also be excluded and thus the model will not necessarily apply in those regions. The use of a risk model as a tool to plan interventions requires further considerations, including the need for high-quality evidence that demonstrates the efficacy of proposed interventions. The generation of high-quality evidence on interventions is needed to achieve beneficial population outcomes informed by the tool. In addition, despite effectively identifying high-risk groups with these tools, it is important to note that accessibility may be a factor preventing groups from benefiting from policies, programs, or interventions. Therefore, in addition to the availability of high-quality evidence on interventions, accessibility is an additional factor that must be considered. Finally, users must assess the impact of potential extreme and rare values on subgroups of risk to ensure they are not overly influential or creating instability.

Conclusions

We anticipate that AvHPoRT will be a valuable addition to the tools used by regional, provincial, and national decision-makers to support ongoing population health management and public health planning.

Abbreviations

ACSC	Ambulatory care sensitive condition
CIHI	Canadian Institutes for Health Information
CCHS	Canadian Community Health Survey
DAD	Discharge Abstract Database
BMI	Body mass index
TRIPOD	Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis
Can-MARG	Canadian Marginalization Index
df	Degrees of freedom
AIC	Akaike information criterion
BIC	Bayesian information criterion

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41512-024-00165-5.

Additional file 1. Summary of health risk behaviour questions and response options from the Canadian Community Health Survey.

Acknowledgements

We are grateful to Philip Cavicchia from Peel Public Health for the feedback used to inform the development of AvHPoRT. The analysis proposed in this protocol will be 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 program, funded by UKRI's Global Challenge Research Fund (Grant Ref: ES/P011055/1). All views expressed in this work are our own.

Authors' contributions

LR conceptualized the study and received study funding. LR, MH, and MO were involved in the study design and protocol drafting and editing. LP, LD, KK, SF, AH, and DM were involved in providing design input and reviewing and revising the protocol. All authors read, provided feedback, and approved the final protocol.

Funding

This work was supported by the Canadian Institutes of Health Research Operating Grant held by LR (FRN 72056684). Laura Rosella is supported by a Canada Research Chair (FRN 72051628).

Availability of data and materials

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 license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Research Data Centre.

Declarations

Ethics approval and consent to participate

This study was approved by the University of Toronto Research Ethics Board (Protocol #37499). This is a register-based study where all data are de-identified and thus there is no ability to directly consent patients. The use of data in this project was authorized under section 45 of Ontario's Personal Health Information Protection Act.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Dalla Lana School of Public Health, University of Toronto, 155 College Street, Health Sciences Building 6th Floor, Toronto, ON M5T 3M7, Canada. [2]Institute for Better Health, Trillium Health Partners, Mississauga, ON, Canada. [3]Laboratory Medicine and Pathobiology, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada. [4]ICES, Toronto, ON M4N 3M5, Canada. [5]PEAK Urban Research Programme, Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK. [6]Department of City & Metropolitan Planning, University of Utah, Salt Lake City, UT, USA. [7]The George Institute for Global Health, Newtown, NSW, Australia. [8]Ottawa Hospital Research Institute, Ottawa, Canada. [9]Statistics Canada, Ottawa, Canada. [10]Department of Family Medicine, University of Ottawa, Ottawa, Canada. [11]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada. [12]Bruyère Research Institute, Ottawa, Canada.

Received: 28 April 2023 Accepted: 15 January 2024
Published online: 06 February 2024

References

1. Billings J, Zeitel L, Lukomnik J, Carey TS, Blank AE, Newman L. Impact of socioeconomic status on hospital use in New York City. Health Aff (Millwood). 1993;12(1):162–73.
2. Billings J, Anderson GM, Newman LS. Recent findings on preventable hospitalizations. Health Aff (Millwood). 1996;15(3):239–49.
3. Canadian Institute for Health Information. Ambulatory care sensitive conditions. Toronto: CIHI; 2018. Available from: http://indicatorlibrary.cihi.ca/display/HSPIL/Ambulatory+Care+Sensitive+Conditions.
4. Victoria State Government. Victorian Health Information Surveillance System: a brief guide to ACSC reports. 2007. Available from: https://www2.health.vic.gov.au/public-health/population-health-systems/health-status-of-victorians/interactive-data-on-the-health-of-victorians/victorian-health-information-surveillance-system.
5. Brown AD, Goldacre MJ, Hicks N, Rourke JT, McMurtry RY, Brown JD, et al. Hospitalization for ambulatory care-sensitive conditions: a method for comparative access and quality studies using routinely collected statistics. Can J Public Health. 2001;92(2):155–9.
6. Busby J, Purdy S, Hollingworth W. How do population, general practice and hospital factors influence ambulatory care sensitive admissions: a cross sectional study. BMC Fam Pract. 2017;18(1):67.
7. Canadian Institute for Health Information. Disparities in primary health care experiences among Canadians with ambulatory care sensitive conditions. Ottawa: Canadian Institute for Health Information; 2012.
8. Wilk P, Ali S, Anderson KK, Clark AF, Cooke M, Frisbee SJ, et al. Geographic variation in preventable hospitalisations across Canada: a cross-sectional study. BMJ Open. 2020;10(5):e037195.
9. Canadian Institute for Health Information. Hospitalization disparities by socio-economic status for males and females. Ottawa: Canadian Institute for Health Information; 2010.
10. Agha MM, Glazier RH, Guttmann A. Relationship between social inequalities and ambulatory care-sensitive hospitalizations persists for up to 9 years among children born in a major Canadian urban center. Ambul Pediatr. 2007;7(3):258–62.
11. Canadian Institute for Health Information. Hospitalization disparities by socio-economic status for males and females. 2010. Available from: https://secure.cihi.ca/free_products/disparities_in_hospitalization_by_sex2010_e.pdf.
12. Falster MO, Jorm LR, Douglas KA, Blyth FM, Elliott RF, Leyland AH. Sociodemographic and health characteristics, rather than primary care supply, are major drivers of geographic variation in preventable hospitalizations in Australia. Med Care. 2015;53(5):436–45.
13. Laditka JN, Laditka SB. Race, ethnicity and hospitalization for six chronic ambulatory care sensitive conditions in the USA. Ethn Health. 2006;11(3):247–63.
14. O'Neil SS, Lake T, Merrill A, Wilson A, Mann DA, Bartnyska LM. Racial disparities in hospitalizations for ambulatory care–sensitive conditions. Am J Prev Med. 2010;38(4):381–8.
15. Pappas G, Hadden WC, Kozak LJ, Fisher GF. Potentially avoidable hospitalizations: inequalities in rates between US socioeconomic groups. Am J Public Health. 1997;87(5):811–6.
16. Wallar LE, Rosella LC. Risk factors for avoidable hospitalizations in Canada using national linked data: a retrospective cohort study. PLoS One. 2020;15(3):e0229465.
17. Tran B, Falster MO, Douglas K, Blyth F, Jorm LR. Smoking and potentially preventable hospitalisation: the benefit of smoking cessation in older ages. Drug Alcohol Depend. 2015;150:85–91.
18. Chew RB, Bryson CL, Au DH, Maciejewski ML, Bradley KA. Are smoking and alcohol misuse associated with subsequent hospitalizations for ambulatory care sensitive conditions? J Behav Health Serv Res. 2011;38(1):3–15.
19. Borda-Olivas A, Fernández-Navarro P, Otero-García L, Sanz-Barbero B. Rurality and avoidable hospitalization in a Spanish region with high population dispersion. Eur J Pub Health. 2012;23(6):946–51.
20. Chen CC, Chen LW, Cheng SH. Rural–urban differences in receiving guideline-recommended diabetes care and experiencing avoidable hospitalizations under a universal coverage health system: evidence from the past decade. Public Health. 2017;151:13–22.
21. Cloutier-Fisher D, Penning MJ, Zheng C, Druyts E-BF. The devil is in the details: trends in avoidable hospitalization rates by geography in British Columbia, 1990–2000. BMC Health Serv Res. 2006;6(1):104.
22. Hale N, Probst J, Robertson A. Rural area deprivation and hospitalizations among children for ambulatory care sensitive conditions. J Community Health. 2016;41(3):451–60.
23. Maria Sanchez SVJHJL, Hui J. CIHI survey: variations in Canadian rates of hospitalization for ambulatory care sensitive conditions. Healthc Q. 2008;11(4):20–2.
24. Weissman JS, Gatsonis C, Epstein AM. Rates of avoidable hospitalization by insurance status in Massachusetts and Maryland. JAMA. 1992;268(17):2388–94.
25. Bocour A, Tria M. Preventable hospitalization rates and neighborhood poverty among New York City residents, 2008–2013. J Urban Health. 2016;93(6):974–83.
26. Roos LL, Walld R, Uhanova J, Bond R. Physician visits, hospitalizations, and socioeconomic status: ambulatory care sensitive conditions in a Canadian setting. Health Serv Res. 2005;40(4):1167–85.
27. Wallar LE, Rosella LC. Individual and neighbourhood socioeconomic status increase risk of avoidable hospitalizations among Canadian adults: a retrospective cohort study of linked population health data. Int J Popul Data Sci. 2020;5(1):1351.
28. Dantas I, Santana R, Sarmento J, Aguiar P. The impact of multiple chronic diseases on hospitalizations for ambulatory care sensitive conditions. BMC Health Serv Res. 2016;16(1):348.
29. Walker RL, Chen G, McAlister FA, Campbell NRC, Hemmelgarn BR, Dixon E, et al. Hospitalization for uncomplicated hypertension: an ambulatory care sensitive condition. Can J Cardiol. 2013;29(11):1462–9.
30. Hossain MM, Laditka JN. Using hospitalization for ambulatory care sensitive conditions to measure access to primary health care: an application of spatial structural equation modeling. Int J Health Geogr. 2009;8:51.
31. Laberge M, Wodchis WP, Barnsley J, Laporte A. Hospitalizations for ambulatory care sensitive conditions across primary care models in Ontario, Canada. Soc Sci Med. 2017;181:24–33.
32. Guttmann A, Shipman SA, Lam K, Goodman DC, Stukel TA. Primary care physician supply and children's health care use, access, and outcomes: findings from Canada. Pediatrics. 2010;125(6):1119–26.
33. Ansari Z, Laditka JN, Laditka SB. Access to health care and hospitalization for ambulatory care sensitive conditions. Med Care Res Rev. 2006;63(6):719–41.
34. Oliver-Baxter J, Bywood P, Erny-Albrecht K. Predictive risk models to identify people with chronic conditions at risk of hospitalisation. Adelaide: PHCRIS Policy Issue Review; 2015.
35. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. JAMA. 2011;306(15):1688–98.
36. Wallace E, Stuart E, Vaughan N, Bennett K, Fahey T, Smith SM. Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. Med Care. 2014;52(8):751–65.
37. Zhou H, Della PR, Roberts P, Goh L, Dhaliwal SS. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. BMJ Open. 2016;6:e011060.
38. Health Quality Ontario, Canadian Institute for Health Information. Early identification of people at risk of hospitalization: Hospital Admission Risk Prediction (HARP) - a new tool for supporting providers and patients. Toronto: Canadian Institute for Health Information; 2013.

39. Rosella L, Kornas K. Putting a population health lens to multimorbidity in Ontario. Healthc Q. 2018;21(3):8–11.
40. Manuel DG, Rosella LC. Commentary: assessing population (baseline) risk is a cornerstone of population health planning—looking forward to address new challenges. Int J Epidemiol. 2010;39(2):380–2.
41. Hewitt AM, Mascari JL, Wagner SL. Population health management: strategies, tools, applications, and outcomes. 1st ed. New York: Springer Publishing Company, LLC; 2021.
42. Peat G, Riley RD, Croft P, Morley KI, Kyzas PA, Moons KG, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. PLoS Med. 2014;11(7):e1001671.
43. Statistics Canada. Canadian Community Health Survey - Annual Component (CCHS). Ottawa: Statistics Canada; 2018. Available from: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=795204.
44. Canadian Institute for Health Information. Data quality documentation, discharge abstract database—multi-year information. 2012. Available from: https://www.cihi.ca/sites/default/files/dad_multi-year_en_0.pdf.
45. Statistics Canada. Canadian community health survey data (2000 to 2011) linked to the discharge abstract database (1999/2000–2012/2013). 2018. Available from: https://www.statcan.gc.ca/eng/rdc/cencchs-dad.
46. Sanmartin C, Decady Y, Trudeau R, Dasylva A, Tjepkema M, Fines P, et al. Linking the Canadian Community Health Survey and the Canadian Mortality Database: an enhanced data source for the study of mortality. Health Rep. 2016;27(12):10–8.
47. Matheson FI, Dunn JR, Smith KL, Moineddin R, Glazier RH. Development of the Canadian Marginalization Index: a new tool for the study of inequality. Can J Public Health. 2012;103(8 Suppl 2):S12–6.
48. Statistics Canada. Illustrated glossary: dissemination area (DA). 2017. Available from: https://www150.statcan.gc.ca/n1/pub/92-195-x/2016001/geo/da-ad/da-ad-eng.htm.
49. Matheson FI, Dunn JR, Smith KL, Moineddin R, Glazier RH. Development of the Canadian Marginalization Index: a new tool for the study of inequality. Can J Public Health. 2012;103(8 Suppl 2):S12–6.
50. Rosella LC, Kornas K, Yao Z, Manuel DG, Bornbaum C, Fransoo R, 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(10):e61–9.
51. Rosella LC, Manuel DG, Burchill C, Stukel TA, Phiat DT. 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(7):613–20.
52. Lebenbaum M, Espin-Garcia O, Li Y, Rosella LC. Development and validation of a population based risk algorithm for obesity: the Obesity Population Risk Tool (OPoRT). PLoS One. 2018;13(1):e0191169.
53. Taljaard M, Tuna M, Bennett C, Perez R, Rosella L, Tu JV, 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(10):e006701.
54. Fisher S, Hsu A, Mojaverian N, Taljaard M, Huyer G, Manuel DG, et al. Dementia Population Risk Tool (DemPoRT): study protocol for a predictive algorithm assessing dementia risk in the community. BMJ Open. 2017;7(10):e018018.
55. Manuel DG, Tuna M, Perez R, Tanuseputro P, Hennessy D, Bennett C, et al. Predicting stroke risk based on health behaviours: development of the Stroke Population Risk Tool (SPoRT). PLoS One. 2015;10(12):e0143342.
56. Ng R, Sutradhar R, Kornas K, Wodchis WP, Sarkar J, Fransoo R, 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(6):e204669.
57. Manuel DG, Perez R, Sanmartin C, Taljaard M, Hennessy D, Wilson K, 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(8):e1002082.
58. Rosella LC, O'Neill M, Fisher S, Hurst M, Diemert L, Kornas K, 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(1):18.
59. Canadian Institute for Health Information. Canadian coding standards for version 2015 ICD-10-CA and CCI. 2015. Available from: https://secure.cihi.ca/free_products/Coding%20standard_EN_web.pdf.
60. Sanmartin CA, Khan S, LHAD research team. Hospitalizations for ambulatory care sensitive conditions (ACSC): the factors that matter. 2011. Available from: https://www150.statcan.gc.ca/n1/en/pub/82-622-x/82-622-x2011007-eng.pdf?st=X-5w86du.
61. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441.
62. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med. 2019;38(7):1276–96.
63. Steyerberg EW. Clinical prediction models. 2nd ed. Switzerland: Springer Nature; 2019.
64. Harrell FE. Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001. p. 45–61. Chapter 3.
65. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.
66. Heus P, Damen J, Pajouheshnia R, Scholten R, Reitsma JB, Collins GS, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. BMC Med. 2018;16(1):120.
67. HMisc package. Available from: http://biostat.mc.vanderbilt.edu/wiki/Main/Hmisc. Accessed 9 Apr 2023.
68. Core Team R. R: a language and environment for statistical computing. 2016.
69. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med. 2015;162(10):735–6.
70. Shields M, Connor Gorber S, Janssen I, Tremblay MS. Bias in self-reported estimates of obesity in Canadian health surveys: an update on correction equations for adults. Health Rep. 2011;22(3):35–45.
71. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res. 2007;16(3):219–42.
72. Buuren SV, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Soft. 2010;45(3):1–67.
73. Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59(10):1092–101.
74. Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. Biom J. 2015;57(4):614–32.
75. Ahmed SE, Hossain S, Doksum KA. LASSO and shrinkage estimation in Weibull censored regression models. J Stat Plan Inference. 2012;142(6):1273–84.
76. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–38.
77. Cook NR. Comment: measures to summarize and compare the predictive capacity of markers. Int J Biostat. 2010;6(1):22, discussion Article 5.
78. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem. 2008;54(1):17–23.
79. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58(5):475–83.
80. Peirson L, Rosella L. Navigating knowledge to action: a conceptual map for facilitating translation of population health risk planning tools into practice. J Contin Educ Health Prof. 2015;35(2):139–47.
81. Rosella LC, Bornbaum C, Kornas K, Lebenbaum M, Peirson L, Fransoo R, et al. Evaluating the process and outcomes of a knowledge translation approach to supporting use of the Diabetes Population Risk Tool (DPoRT) in public health practice. Can J Program Eval. 2018;33(1):21–48.
82. Rotermann M. Evaluation of the coverage of linked Canadian Community Health Survey and hospital inpatient records. Health Rep. 2009;20(1):45–51.
83. Gupta N, Crouse DL. Social disparities in the risk of potentially avoidable hospitalization for diabetes mellitus: an analysis with linked census and hospital data. Can Stud Popul. 2019;46(2):145–59.

Publisher's Note