BMJ Open

Cardiovascular Disease Population Risk Tool (CVDPoRT): predictive algorithm for assessing CVD risk in the community setting. A study protocol

Monica Taljaard,[1,2] Meltem Tuna,[1,3] Carol Bennett,[1,3] Richard Perez,[1,3] Laura Rosella,[3,4,5] Jack V Tu,[3,6,7] Claudia Sanmartin,[8] Deirdre Hennessy,[8] Peter Tanuseputro,[1,3,10] Michael Lebenbaum,[3] Douglas G Manuel[1,2,3,8,9,10]

To cite: 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. doi:10.1136/bmjopen-2014-006701

▸ Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2014-006701).

Received 22 September 2014
Accepted 2 October 2014

For numbered affiliations see end of article.

Correspondence to
Dr Monica Taljaard;
mtaljaard@ohri.ca

ABSTRACT

Introduction: Recent publications have called for substantial improvements in the design, conduct, analysis and reporting of prediction models. Publication of study protocols, with prespecification of key aspects of the analysis plan, can help to improve transparency, increase quality and protect against increased type I error. Valid population-based risk algorithms are essential for population health planning and policy decision-making. The purpose of this study is to develop, evaluate and apply cardiovascular disease (CVD) risk algorithms for the population setting.

Methods and analysis: The Ontario sample of the Canadian Community Health Survey (2001, 2003, 2005; 77 251 respondents) will be used to assess risk factors focusing on health behaviours (physical activity, diet, smoking and alcohol use). Incident CVD outcomes will be assessed through linkage to administrative healthcare databases (619 886 person-years of follow-up until 31 December 2011). Sociodemographic factors (age, sex, immigrant status, education) and mediating factors such as presence of diabetes and hypertension will be included as predictors. Algorithms will be developed using competing risks survival analysis. The analysis plan adheres to published recommendations for the development of valid prediction models to limit the risk of overfitting and improve the quality of predictions. Key considerations are fully prespecifying the predictor variables; appropriate handling of missing data; use of flexible functions for continuous predictors; and avoiding data-driven variable selection procedures. The 2007 and 2009 surveys (approximately 50 000 respondents) will be used for validation. Calibration will be assessed overall and in predefined subgroups of importance to clinicians and policymakers.

Ethics and dissemination: This study has been approved by the Ottawa Health Science Network Research Ethics Board. The findings will be disseminated through professional and scientific conferences, and in peer-reviewed journals. The algorithm will be accessible electronically for population and individual uses.

Trial registration number: ClinicalTrials.gov NCT02267447.

Strengths and limitations of this study

- The Cardiovascular Disease Population Risk Tool (CVDPoRT) will use data on major health behavioural risk factors from large population-based community health surveys individually linked to routinely-collected health administrative data in Ontario, Canada, to develop and validate a population-based risk algorithm for CVD.
- CVDPoRT will improve the ability to answer key policy questions with respect to the future burden of CVD in Canada, the contribution of major health behaviours to the population burden of CVD, the preventive benefit of achieving health behaviour goals and strategies to reduce inequities through improvements in health behaviours.
- The analysis plan adheres to published recommendations for the development of valid risk prediction models to limit the risk of overfitting and improve the quality of predictions.
- Although a rigorous approach will be used to develop the model, including internal and external validation, stronger forms of validation may be required: future validation studies should include application in different geographic locations, and fully independent validation by independent investigators using alternative measurement of these risk factors in different population settings.
- The model development will focus on maximising predictive accuracy and as such, will not consider the causal and mediator effects of the predictive variables.

INTRODUCTION

Disease risk algorithms for the population setting

Numerous prognostic models have been developed to predict the risk of future disease for individual patients in clinical settings. Population-based prognostic models are less common, but are essential for

population health planning and policy decision-making. Unlike clinical models, they are usually derived using population data and may utilise self-reported risk factors that do not require laboratory or clinical measurement. The Cardiovascular Disease Population Risk Tool (CVDPoRT) will use data on major health behavioural risk factors (smoking, diet, physical activity, and alcohol use) from a large population-based community health survey individually linked to routinely-collected health administrative data in Ontario, Canada, to develop and validate a population-based risk algorithm for CVD. Once validated, CVDPoRT will improve the ability to answer key policy questions with respect to the future burden of CVD in Canada, the contribution of major health behaviours to the population burden of CVD, the preventive benefit of achieving health behaviour goals, and strategies to reduce inequities through improvements in health behaviours.

Countries with clinical guidelines for CVD generally recommend patient risk assessment and stratification using multivariable risk algorithms, such as the Framingham risk tool.[1–3] Improving *population* health risk assessment has been identified as a priority in Canada.[4 5] Clinical risk algorithms, such as Framingham, are challenging to adapt for population health planning because they require clinical measures such as blood pressure and lipid levels. Although it can be more challenging to develop prediction models that have acceptable discrimination and calibration without the use of clinical measures, previous studies suggest that approximately 50% of CVD may be related to health behaviours[6 7]; moreover, we have previously demonstrated that disease risks can be accurately assessed for population uses using only self-reported risk factors.[8–10] There are several advantages to developing population-based prediction models without clinical measures. First, surveys that assess only self-reported risk factors are usually much larger than those that include clinical measures. Second, since population-based health surveys are now being conducted in over 100 countries, such algorithms have a broader scope of potential application. Third, population-based algorithms allow for estimation of population-level disease risk. Fourth, the inclusion of self-reported risk factors and health behaviours complements existing clinical risk algorithms whose focus is biophysical measures such as lipids and hypertension. Finally, self-reported risk factors are easily ascertained by individuals, which facilitates implementation of internet-based risk calculators in community settings.

It is recognised that self-reported risk factors may introduce greater measurement error than biophysical or clinical measures, and that this may adversely affect the performance of prognostic models. Whereas demographic characteristics such as education and some behaviours such as smoking are usually measured with little error, other risk factors such as diet, physical activity, body mass index (BMI), and especially alcohol consumption, may have more substantial bias.[11] These errors may be minimised through the use of a comprehensive set of sociodemographic and behavioural risk factors that are commonly ascertained in population health surveys, using reliable methods with well-established exposure questions and limited rates of missing data. Moreover, the effect of omission of biophysical measures (eg, measured blood pressure) may be minimised by including variables that are correlated with such measures (eg, blood pressure medication).

Methodological issues in prediction model research

Prediction models are more likely to be reliable and useful in practice when they are developed using a large, high-quality data set; based on a study protocol with a sound statistical analysis plan; and validated in independent data sets.[12] Among thousands of clinical prediction rules published in the past decades, many have been shown to have serious methodological shortcomings.[13] A review of 83 clinical prediction models in acute stroke, for example, found serious deficiencies in statistical methods in almost all of the studies; in addition, none of the studies had been adequately validated.[14] A series of recent publications have called for substantial improvements in the design, conduct, analysis and reporting of prognostic studies.[12 15–17] Several threats to validity have been identified, including inadequate sample sizes, data-driven or arbitrary categorisation of continuous predictors, inadequate statistical modelling of non-linear relationships, inappropriate handling of missing data, and failure to check model assumptions. Statistical overfitting is a particular concern in the development of prognostic models; it results when a model is fitted with too many parameters given the amount of information in the data. In such circumstances, the predictive ability of the model will be overstated and it is likely to perform poorly in different settings.[18] When overfitting is present, some of the associations in the model may be spurious, reflecting increased type I error. The use of tests of association for selecting predictor variables, data-driven categorisation or specification of functional form of association with predictors, and stepwise variable selection procedures can increase the risk of type I error.

Before any prognostic model might be adopted in practice it is necessary to show that it provides valid predictions outside the specific context of the sample used to derive the model.[19] A recent review of 71 published clinical prediction models in high-impact journals found that only one study included external validation during development, and two recalibrated algorithms after publication.[13] Furthermore, none of the 71 algorithms examined calibration for target populations beyond arbitrary risk categories such as deciles of predicted risk. Most focused on discrimination, rather than calibration. This is likely a reflection of the emphasis on identifying high-risk patients during clinical decision-making[20]; for population uses, however, several authors have emphasised the importance of examining calibration in

important target populations, particularly when these populations and exposures were not included during algorithm development.[21–24] This reflects the intended application of population-based risk algorithms, which includes assessing resource allocation, equity issues, impact of population-wide prevention strategies and disease burden for different levels of exposure.

Transparency in prediction model research

There are several benefits to publishing study protocols: it may improve study quality through peer review; it allows readers to compare what was originally intended and what was actually done, thus preventing both 'data dredging', post-hoc revisions of study aims, and selective reporting; it enables funders and researchers to see what studies are underway and hence reduce duplication of research effort; it enhances credibility of the research by allowing others to replicate the study; and it allows easier identification of and access to details of the study. Peat *et al*[25] have stressed the importance of predefining the key aspects of a prognostic study—yet, it seems that most prognostic studies are conducted without a study protocol, with analysis plans being developed during or after data collection. Although it is recognised that a prognosis research protocol cannot be a rigid blueprint and that it is neither possible nor desirable to pre-specify all analyses,[25] the development of CVDPoRT is especially amenable to prespecification because risk factors for CVD have been well studied and many have known relationships to CVD. Given the goal of generalising to other population-based settings, it is particularly important to avoid overfitting. We are presenting our study protocol to improve transparency and protect against bias. Our protocol adheres to a recommended checklist of items to include in protocols for prognostic studies.[25]

Objectives

The objective of this study is to develop and validate CVD risk prediction models for the population setting using self-reported risk factors with a focus on major health behaviours. We will use the Ontario sample of the Canadian Community Health Survey (CCHS) individually linked to routinely-collected data to ascertain CVD incident events. Separate models will be derived for men and women.

METHODS AND ANALYSIS

Outcomes

The primary outcome of interest is a major CVD event, ascertained using validated diagnostic codes and criteria as presented in table 1. Additional prediction models will be derived for secondary outcomes of interest, defined in table 2.

Design

CVDPoRT will be derived and validated using secondary data. The derivation cohort will be eligible respondents to the combined 2001, 2003 and 2005 Canadian Community Health Surveys (CCHS cycles 1.1, 2.1 and 3.1), conducted by Statistics Canada. The CCHS surveys use a multistage stratified cluster design that represents approximately 98% of the Canadian population aged 12 years and above, and attains an average response rate of 80.5%. The surveys are conducted through telephone and in-person interviews and all responses are self-reported. The details of the survey methodology have been previously published.[26] All self-reported risk factors of interest will be obtained from the CCHS. To ascertain CVD events, the survey respondents will be individually linked to two population-based databases: hospitalisation records from the Canadian Institute for Health Information Discharge Abstract Database, and vital statistics. Secondary outcomes will also require linkage to the Ontario Health Insurance Plan database. Respondents will be followed until the earliest of: incident event, death (defined as a competing risk), loss to follow-up (defined as loss of healthcare eligibility), or end of study (31 December 2011 or most recent year available). The validation cohort will consist of respondents to the 2007 and 2009 surveys, similarly linked to ascertain outcomes. Owing to the known challenges of using survey weights in regression models,[27] including difficulties in obtaining correct estimates for SE, and complexity of modelling procedures and interpretation of results, no survey weights will be incorporated in the development of CVDPoRT.

Table 1 Diagnostic codes for CVD main events

Definition	ICD-9	ICD-10
Hospitalisation (main diagnosis only)		
Acute myocardial infarction	410	I21
	410	I22
Stroke	430	I60
	431	I61
	434	I63 excluding I63.6
	436	I64
	362.3	H341
Death (vital statistics)		
Ischemic heart disease death	410–414, 429.2	I20–I25
Stroke death	430–434, 436–438	I60–I69

CVD, cardiovascular disease; ICD, The International Classification of Diseases, Ninth Revision.

Eligibility criteria

Respondents will be excluded if they were not eligible for Ontario's universal health insurance programme, were pregnant, self-reported a history of heart disease or stroke, or were younger than age 20 at the time of survey administration. If a respondent was included in

Table 2 Diagnostic codes for secondary outcomes

Outcome	Definition	ICD-9	ICD-10
(1) CVD—total	Major cardiovascular disease	CVD main events as defined in table 1	CVD main events as defined in table 1
	Transient ischemic attack	435	G45 (excluding G454)
	Congestive heart failure	428	I50
	Other acute coronary syndrome	411, 413	I20, I23.82, I24
	Peripheral vascular disease (amputation and bypass)	**CCP codes**	**CCI codes**
	Leg amputation*	96.14, 96.15	1VQ93, 1VC93, 1VG93
	Foot or toe amputation*	96.11, 96.12, 96.13	1WM93, 1WL93, 1WA93, 1WE93, 1WJ93
	Arterial bypass surgery†	51.25, 51.29	1KG76
	Percutanous transluminal angioplasty†	50.18	1KG50, 1KG57,1KG76, 1KG35HAC1, 1KG35HHC1
(2) Major coronary artery disease	Acute myocardial infarct	410	I21, I22
(3) Coronary artery disease—total	Acute myocardial infarct	410	I21, I22
	Other acute coronary syndrome	411, 413	I20, I23.82, I24
(4) Stroke—major	Stroke hospitalisation or death	Hospital—430, 434, 436, 362.3 Death—430–434, 436–438	I60, I61, I63 (excluding I63.6), I64, H341 I60—I69
(5) Stroke—minor	Stroke—major and stroke, hospitalised TIA, stroke or TIA diagnosed in the outpatient setting	Same as stroke—major Stroke in the community setting as ascertained by Tu *et al*[35]	Same as stroke—major Stroke in the community setting as ascertained by Tu *et al*[35]
Sensitivity testing (inclusion of less commonly used diagnostic codes)			
	Acute stroke	362.3	I65
	Stroke/TIA	437.1, 437.9, 438	I67.8, I67.9, I69

*Exclude all upper leg or foot amputations if in conjunction with [ICD9: 170, 171, 213, 730, 740–759, 800–900, 901–904, 940–950 ICD10: C40, C41, C46.1, C47, C49, D160, M46.2, M46.2, M86, M87, M89.6, M90.0-M90.5, Q00, Q38-Q40, S02.0, S09.0, S04.0, S15, S25, S25, T26].
†Exclude all records with a diagnosis code of aneurysm [ICD9: 4141, 441, 442, ICD10: I67.1, I71, I72, I60, 177.0, 179.0, Q codes].
CCI, Canadian Classification of Health Interventions; CCP, Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures; CVD, cardiovascular disease; ICD, The International Classification of Diseases, Ninth Revision; TIA, transient ischemic attack.

more than one CCHS cycle, only their earliest survey response will be used. The same exclusion criteria will be applied to the respondents in the validation cohort.

Sample size

The derivation cohort consists of 77 251 respondents and 619 886 person-years of follow-up until 31 December 2011; the validation cohort will consist of approximately 50 000 respondents and 150 000 person-years of follow-up. The number of events until 31 December 2011 in the derivation cohort is 1131 for men and 1102 for women; in the validation cohort we expect approximately 250 events for men and 250 for women. Harrell[18] describes sample size requirements for prediction models. For time to event outcomes, the number of participants experiencing the event must exceed 10 times the number of degrees of freedom, where the number of degrees of freedom includes the number of predictors screened for association with the outcome, all dummy variables, non-linear terms and interactions. For CVDPoRT, the target number of total regression degrees of freedom is less than 110. The minimum sample size requirement for external validation studies is 100 events and 100 non-events.

Analysis plan

We closely followed guidelines by Harrell[18] and Steyerberg[28] in the development of our analysis plan, which was constructed after accessing the derivation data set, but prior to any model fitting or any descriptive analyses involving the exposure-outcome associations. Key considerations in our approach were fully prespecifying the predictor variables, use of flexible functions for continuous predictors, and preserving statistical properties by avoiding data-driven variable selection procedures. Analyses will be conducted using Harrell's HMisc[29] package of functions in R[30] as well as SAS V.9.3.

Identification of predictors

Identification of predictor variables was based on reviewing the available data collected across all cycles in the CCHS together with subject-matter expertise, and was informed by our previous work in developing models for diabetes, stroke, and life expectancy.[9 10] Questionnaires used in the CCHS are available elsewhere.[31] The following categories of predictors were considered: sociodemographic, health behaviours and morbidities. Some variables needed to be constructed from multiple items in the survey

questionnaire. Variables with more than 20% missing values were excluded from consideration. Variables with narrow distributions or insufficient variation were excluded. Obvious cases of redundancy (eg, alternative definitions of the same underlying behaviour) were ruled out. A formal check of multicollinearity was carried out using a variable clustering algorithm.[18] A total of 22 predictor variables were finally identified: 7 sociodemographic, 11 behavioural, and 3 disease risk factors, with 1 design variable. Education —rather than individual income—was selected as a predictor due to several concerns associated with income, including lack of generalisability, measurement error, stability overtime and substantial missingness. An indicator variable for immigration status together with fraction of life lived in Canada was used to account for recent and non-recent immigrants. Indicator variables for smoking status were created to allow inclusion of smoking pack-years as a continuous predictor. The model will additionally include interactions between age and: smoking, alcohol, diet, physical activity, BMI, diabetes, and hypertension, as the effect of these risk factors on CVD are expected to vary with age. Detailed definitions and measurement of these variables are presented in table 3.

Data cleaning and coding of predictors

Data cleaning and coding will proceed without examining outcome-risk factor associations. Coding of variables will focus on minimising the loss of predictive information by avoiding categorisation. Continuous variables will be inspected using boxplots and descriptive statistics to

Table 3 Prespecified predictor variables for CVDPoRT with initial degrees of freedom (df) allocation

Variable	Scale	Valid range/levels	df
Demographic			
Age	Continuous	20 to 105	4
Sex	Dichotomous	Male, female	NA
Health behaviours			
Pack years of smoking	Continuous	0–310	2
Smoking status	Categorical	Non-smoker; current smoker; former smoker quit <5 years ago; former smoker quit >5 years ago	3
Alcohol consumption (number of drinks last week)	Continuous	0–170	2
Former drinker	Dichotomous	Yes, no	1
Consumption of fruit, salad, carrot and other vegetables (average daily frequency)	Continuous	0.0 to 80.0	2
Potato consumption (average daily frequency)	Continuous	0.0 to 20.0	2
Juice consumption (average daily frequency)	Continuous	0.0 to 20.0	2
Leisure physical activity (average daily meets (kcal/kg/day))	Continuous	0.0 to 35.0	2
Self-perceived stress	Ordinal	Not at all stressful; not very stressful; a bit stressful; quite a bit stressful; extremely stressful	4
Sense of belonging to local community	Ordinal	Very strong; somewhat strong; somewhat weak; very weak	3
Body mass index	Continuous	8.8–120	2
Sociodemographic			
Ethnicity	Categorical	Caucasian; African–American; Chinese; Aboriginal; Japanese/Korean/South East Asian/Filipino; Other/ Multiple origin/ Unknown/ Latin American; South Asian/Arab/West Asian	6
Immigrant	Dichotomous	Yes, no	1
% of life lived in Canada	Continuous	0–100%	2
Education	Categorical	Less than secondary school; secondary school graduation; some postsecondary; postsecondary graduation	3
Neighbourhood social and material deprivation	Ordinal	Pampalon's deprivation index[36]: low (1st or 2nd quintile); high (4th or 5th quintile); moderate (all others)	1
Diseases			
Diabetes	Dichotomous	Yes, no	1
High BP	Dichotomous	Yes, no	1
High BP medication	Dichotomous	Yes, no	1
Design			
Survey year	Categorical	2001, 2003, 2005	2

BP, blood pressure; CVDPoRT, Cardiovascular Disease Population Risk Tool; NA, not applicable.

determine values outside a plausible range. Values that are clearly erroneous will be corrected, where possible, or otherwise set to missing. Truncation to the 99.5th centile or where the data density ends will be considered for continuous risk factors with highly skewed distributions (eg, smoking pack-years, diet, alcohol consumption, physical activity) based on inspection of histograms and boxplots. To avoid instability in the regression analyses, frequency distributions for categorical predictors will be examined and categories with small numbers of respondents will be combined.

Missing data

Traditional complete case analyses suffer from inefficiency, selection bias and other limitations.[28] We will use multiple imputation to impute missing values on all predictors, using the 'aregImpute' function in the HMisc library. This procedure simultaneously imputes missing values while determining optimal transformations among all imputation variables. Predictive mean matching is used to replace missing values with random draws of observed values from participants with the nearest predicted values. The imputation model will consist of the full list of predictor variables, along with time to event and censoring variables, as well as auxiliary variables—variables that are not predictors, but may nevertheless be useful in generating imputed values, for example, income and self-perceived health. We will generate five multiple imputation data sets. The final model will be estimated separately for each completed data set and the results combined using the rules developed by Rubin and Schenker[32] to account for imputation uncertainty.

Model specification

Using the approach described by Harrell,[18 29] we will fit an initial main effects model that includes an initial degree of freedom allocation for each predictor. We will then decide how to allocate final numbers of degrees of freedom to individual predictors based on a partial test of association with the outcome. Decisions on initial degree of freedom allocations will be informed by the anticipated importance of each predictor and any known dose–response relationships with CVD (eg, known "U" or "J" shaped relationships for alcohol and BMI). Continuous predictors will be flexibly modelled using restricted cubic splines, that is, piecewise cubic functions that are smooth at the knots and restricted to be linear in the tails. The knots will be placed at fixed quantiles of the distribution: in particular, at the 5th, 27.5th, 50th, 72.5th and 95th centiles. Ordinal variables with few categories will be specified as either linear terms, or as categorical if the expected association is more complex than linear. Interactions will be restricted to linear terms. The initial model specification, presented in table 3, includes a total of 61 degrees of freedom (47 main, 14 interaction), compared to a possible maximum of 110. Partial association χ^2 statistics for each predictor variable minus their degrees of freedom (to level the playing field among predictors with varying degrees of freedom) will be plotted in descending order. Variables with higher predictive potential will be allocated more degrees of freedom, but predictors with lower predictive potential will be modelled as simple linear terms or recoded by combining infrequent categories. As described by Harrell,[18 29] this process of model specification does not increase the type I error because predictors will be retained in the full model regardless of their strength of association, tests of nonlinearity will not be revealed to the analyst and combining categories may include collapsing the most disparate categories as they will also be blinded to the observed rates of events per category.

Model estimation

The initial model will be estimated using competing risks Cox proportional hazards regression with death from a non-CVD cause considered a competing risk; alternative model specifications may need to be considered after assessing validity of model assumptions. All continuous predictors will be centred about their means. A key assumption of the Cox model, that is, that the effect of predictors is constant in time, will be assessed using plots of raw and smoothed scaled Schoenfeld residuals versus time for each predictor. To address serious violations of this assumption, interaction terms between the predictor and log-time will be considered. Influence will be assessed by plotting scaled dfβ residuals for each covariate. Although the risk of overfitting will be minimal due to prespecification of our model and large sample size, we will nevertheless assess the need to adjust for overfitting. The degree of overfitting (shrinkage) in the model will be estimated using the heuristic shrinkage estimator (based on the log likelihood ratio χ^2 statistic for the full model).[33] If shrinkage is <0.90, adjustment for overfitting will be required, as described below.

Assessment of model performance

Steyerberg[28] distinguishes between apparent, internally validated and externally validated model performance. 'Internally validated performance' corrects for optimism in the apparent performance to yield approximately unbiased estimates of future model performance. Performance in the derivation and validation cohorts will be assessed and reported using overall measures of predictive accuracy, discrimination (ability to differentiate between high-risk and low-risk individuals), and calibration (agreement between predicted and observed risk). All model performance measures will be calculated using the first of the multiply imputed data sets. Nagelkerke's R^2 and the Brier score will be calculated as overall measures of accuracy. Discrimination will be assessed using Harrell's overall concordance statistic, with 95% CIs estimated using bootstrap samples. Internally validated performance measures will be

obtained using 200 bootstrap samples, using the procedure described by Steyerberg.[28]

Steyerberg[28] and Cook[21 22] suggest that calibration should receive more attention when evaluating prediction models, and that assessment of recalibration tests and calibration slopes should be used routinely. We will emphasise visualisation of model performance using plots, rather than formal statistical testing: significance of traditional Hosmer-Lemeshow goodness of fit tests, for example, may reflect large sample sizes rather than true miscalibration. Thus, we will create calibration plots at fixed time points by comparing mean predicted probabilities with Kaplan-Meier estimates of observed rates stratified by intervals of predicted risk. The calibration slope will be estimated by including the linear predictor as a single term in the model fitted to the validation cohort. Deviation from a slope of one will be tested using a Wald or likelihood ratio test. The calibration slope reflects the combined effect of overfitting to the derivation data as well as true differences in effects of predictors.

Subgroup validation will be implemented as a conceptually easy check of calibration. This entails comparing observed and predicted risks within predefined subgroups of importance to clinicians and policymakers, for example, defined by age groups, behavioural risk exposure categories, health regions, sociodemographic groups, hypertension status and diabetes status. Explicit criteria for clinically or policy relevant standards of calibration will be established, for example, <20% difference between observed and predicted estimates for categories with prevalence higher than 5%.

Estimation of the final model

Prespecification of predictors has advantages in limiting the risks of overfitting and spurious statistical significance, but may result in a final model that is overly complex and difficult to interpret. It may be possible to derive a more parsimonious model that retains most of the prognostic information, and that performs as well or better than the full model, without increasing the type I error rate.[18 34] We will use the stepdown procedure described by Ambler *et al*[34] to identify a more parsimonious model. This procedure involves deleting variables to a desired degree of accuracy based on contribution to model R^2. We will compare the reduced and full models using internal bootstrap validation, with appropriate penalisation for the variable selection. The final model (either the full model or its approximation) will be selected based on comparing calibration slopes, calibration in subgroups of policy importance, and overall measures of predictive accuracy.

To maximise duration of follow-up, the final regression coefficients will be estimated using the combined data from the derivation and validation cohorts with outcome events updated to reflect the most recent years available. If relevant differences are found between the derivation and validation cohorts, a cohort-specific intercept and/or interaction term may be included in the final model; otherwise, it will maintain the same risk factors and form as the derivation model.

Model presentation

Results will be presented for the derivation, validation and combined cohorts. Given the anticipated complexity of the final regression model, the usual presentation of a regression model showing estimated HRs and 95% CIs is less meaningful. To allow interpretation of the estimated effect of each predictor, the model will be summarised using plots of the shape of the effect of each predictor, as well as Wald χ^2 statistics, penalised for degrees of freedom. As predictions are of primary interest, presentation will take the form of a regression formula, which will serve as the basis for web-based implementation.

Secondary outcome analyses

Prediction models for secondary outcomes will be derived separately with a new development process, but maintaining the same risk factors and model specification as for the primary outcome. Sensitivity analyses will be carried out using less commonly used diagnostic codes (see table 2).

Analyses beyond initial model development

We also plan to also validate the algorithms using the national sample of the individually-linked CCHS 1.1, 2.1, 3.1, when these data become available. We will conduct further analyses exploring the predictive ability of novel risk factors that were not previously included (eg, food insecurity), as well as risk factors that were not ascertained in all CCHS cycles (eg, active transportation, workplace stress, depression and anxiety, cholesterol therapy). Risk factors that can be ascertained through linkages of additional data sources and similar cohorts (eg, area-based measures of built environment, air pollution, detailed dietary consumption, lipid levels, glucose levels, measured blood pressure) will additionally be explored. Diet will be examined using a previously developed diet quality measure (Perez measure[10]) that showed good predictive performance for both all-cause mortality and all-cause hospitalisation; it uses six diet exposure measures with weighting for carrot, potato and fruit juice consumption. These exploratory risk factors will not be included in CVDPoRT, but will be considered in future updates of the model.

ETHICS AND DISSEMINATION

A project advisory committee has been created to ensure that the risk algorithm development meets the needs of knowledge users. The committee has been involved from the beginning of the study and worked with the study team to rank candidate predictors for inclusion based on policy importance and scientific importance. It will advise on the identification of important target populations, and establish minimal

policy important differences for calibration studies. Results from CVDPoRT will be submitted for publication in peer-reviewed journals and presentation at scientific meetings. If appropriate for individual use, we will create a web-based CVD calculator. Although CVDPoRT emphasises population risk prediction, our experience has shown that individual calculators are an effective engagement and translation tool for both the general public and knowledge users.

CONCLUSION

To the best of our knowledge, CVDPoRT will be the first population-based risk prediction algorithm for CVD. Although a rigorous approach will be used to develop the model, including internal and external validation, stronger forms of validation may be required. Future validation studies should include application in different geographic locations, and fully independent validation by independent investigators using alternative measurement of these risk factors in different population settings.

Author affiliations
[1]Ottawa Hospital Research Institute, Ottawa, Ontario, Canada
[2]Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada
[3]Institute for Clinical Evaluative Sciences, Ottawa and Toronto, Ontario, Canada
[4]Public Health Ontario, Toronto, Ontario, Canada
[5]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[6]Sunnybrook Schulich Heart Centre, University of Toronto, Toronto, Ontario, Canada
[7]Institute of Health Policy, Management, and Evaluation, University of Toronto, Toronto, Ontario, Canada
[8]Health Analysis Division, Statistics Canada, Ottawa, Ontario, Canada
[9]Department of Family Medicine, University of Ottawa, Ottawa, Ontario, Canada
[10]Bruyère Research Institute, Ottawa, Ontario, Canada

Contributors DGM is the principal investigator of the study, is responsible for the conception of the project, led the grant application, led the design and statistical analysis plan, critically reviewed the manuscript, and approved the final version. MTu contributed to the grant application, contributed to the design and detailed statistical analysis plan, cleaned and analysed the data, critically reviewed the manuscript, and approved the final version. MTa contributed to the grant application, contributed to the design and detailed statistical analysis plan, drafted this manuscript, critically reviewed the manuscript, and approved the final version. CB and RP contributed to the grant application, contributed to the design and detailed statistical analysis plan, critically reviewed the manuscript, and approved the final version. LR, JVT, CS, DH, PT, and ML contributed to the grant application, reviewed and commented on the design and statistical analysis plan, critically reviewed the manuscript, and approved the final version.

Funding This work was supported by the Canadian Institutes of Health Research operating grant FRN: 133550 and the Cardiovascular Health in Ambulatory Research Team (CANHEART) from the Canadian Institutes of Health Research (TCA 118349). JVT holds a Tier 1 Canada Research Council Chair in Health Services Research and a Career Investigator award from the Heart and Stroke Foundation.

Competing interests None.

Ethics approval Research ethics approval has been obtained from the Ottawa Health Science Network Research Ethics Board.

Provenance and peer review Not commissioned; peer reviewed for ethical and funding approval prior to submission.

REFERENCES

1. Ferket BS, Colkesen EB, Visser JJ, *et al*. Systematic review of guidelines on cardiovascular risk assessment: which recommendations should clinicians follow for a cardiovascular health check? *Arch Intern Med* 2010;170:27–40.
2. Manuel DG. The effectiveness of national guidelines for preventing cardiovascular disease: integrating effectiveness concepts and evaluating guidelines' use in the real world. *Curr Opin Lipidol* 2010;21:359–65.
3. Anderson TJ, Gregoire J, Hegele RA, *et al*. 2012 update of the Canadian Cardiovascular Society guidelines for the diagnosis and treatment of dyslipidemia for the prevention of cardiovascular disease in the adult. *Can J Cardiol* 2013;29:151–67.
4. Smith ER. The Canadian heart health strategy and action plan. *Can J Cardiol* 2009;25:451.
5. Manuel DG, Luo W, Ugnat AM, *et al*. Cause-deleted health-adjusted life expectancy of Canadians with selected chronic conditions. *Chronic Dis Can* 2003;24:108–15.
6. Ford ES, Zhao G, Tsai J, *et al*. Low-risk lifestyle behaviors and all-cause mortality: findings from the National Health and Nutrition Examination Survey III mortality study. *Am J Public Health* 2011;101:1922.
7. Yusuf S, Hawken S, Ounpuu S, *et al*. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* 2004;364:937–52.
8. Manuel DG, Rosella LC, Hennessy D, *et al*. Predictive risk algorithms in a population setting: an overview. *J Epidemiol Community Health* 2012;66:859–65.
9. Rosella LC, Manuel DG, Burchill C, *et al*. A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). *J Epidemiol Community Health* 2011;65:613–20.
10. Manuel DG, Perez R, Bennett C, *et al*. *Seven more years: the impact of smoking, alcohol, diet, physical activity and stress on health and life expectancy in Ontario*. An ICES/PHO Report. Toronto: Institute for Clinical Evaluative Sciences and Public Health Ontario, 2012.
11. Garriguet D, Colley RC. A comparison of self-reported leisure-time physical activity and measured moderate-to-vigorous physical activity in adolescents and adults. *Health Rep* 2014;25:3–11.
12. Steyerberg ET, Moons KGM, van der Windt DA, *et al*. Prognosis research strategy (PROGRESS) 3: Prognostic model research. *PLoS Med* 2013;10:e1001381.
13. Bouwmeester W, Zuithoff NP, Mallett S, *et al*. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9:1–12.
14. Counsell C, Dennis M. Systematic review of prognostic models in patients with acute stroke. *Cerebrovasc Dis* 2001;12:159–70.
15. Hemingway H, Croft P, Perel P, *et al*. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ* 2013;346:e5595.
16. Riley RD, Hayden JA, Steyerberg ET, *et al*. Prognosis research strategy (PROGRESS) 2: prognostic factor research. *PLoS Med* 2013;10:e1001380.
17. Hingorani A, van der Windt DA, Riley RD, *et al*. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ* 2013;346:e5793.
18. Harrell FE. *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis*. New York: Springer 2001.
19. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19:453–73.
20. Diamond GA. What price perfection? Calibration and discrimination of clinical prediction models. *J Clin Epidemiol* 1992;45:85–9.
21. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem* 2008;54:17–23.

22. Cook NR. Comment: measures to summarize and compare the predictive capacity of markers. *Int J Biostat* 2010;6:Article 22; discussion Article 25.
23. Manuel DG, Lim J, Tanuseputro P, *et al*. Revisiting rose: strategies for reducing coronary heart disease. *BMJ* 2006;332:659–62.
24. Manuel DG, Kwong K, Tanuseputro P, *et al*. Effectiveness and efficiency of different guidelines on statin treatment for preventing deaths from coronary heart disease: modelling study. *BMJ* 2006;332:1419.
25. Peat G, Riley RD, Croft P, *et al*. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. *PLoS Med* 2014;11:e1001671.
26. Beland Y. Canadian community health survey—methodological overview. *Health Rep* 2002;13:9–14.
27. Gelman A. Struggles with survey weighting and regression modeling. *Stat Sci* 2007;22:153–64.
28. Steyerberg EW. *Clinical prediction models*. New York: Springer, 2009.
29. HMisc package. http://biostat.mc.vanderbilt.edu/wiki/Main/Hmisc (accessed Sep 2014).
30. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0. http://www.R-project.org
31. Canadian Community Health Survey (CCHS). http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SurvId=1630&InstaId=3359&SDDS=3226. (accessed Sep 2014).
32. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. *Stat Med* 1991;10:585–98.
33. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. *Stat Med* 1990;9:1303–25.
34. Ambler G, Brady AR, Royston P. Simplifying a prognostic model: a simulation study based on clinical data. *Stat Med* 2002;21:3803–22.
35. Tu K, Wang M, Young J, *et al*. Validity of administrative data for identifying patients who have had a stroke or transient ischemic attack using EMRALD as a reference standard. *Can J Cardiol* 2013;29:1388–94.
36. A deprivation index for health planning in Canada. http://www.phac-aspc.gc.ca/publicat/cdic-mcbc/29–4/ar_05-eng.php (accessed Sep 2014).