BMJ Open

BMJ Open

Dementia Population Risk Tool (DemPoRT): Study Protocol for a Predictive Algorithm Assessing Dementia Risk in the Community

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018018
Article Type:	Protocol
Date Submitted by the Author:	31-May-2017
Complete List of Authors:	Fisher, Stacey; Ottawa Hospital Research Institute; University of Ottawa, Department of Epidemiology and Community Medicine Hsu, Amy; Ottawa Hospital Research Institute; Institute of Clinical Evaluative Sciences Mojaverian, Nassim; Institute of Clinical Evaluative Sciences Taljaard, Monica; Ottawa Hospital Research Institute, Clinical Epidemiology Program; University of Ottawa, Department of Epidemiology and Community Medicine Huyer, Greg; Ottawa Hospital Research Institute; University of Ottawa, Department of Epidemiology and Community Medicine Manuel, Douglas; Ottawa Hospital Research Institute, Clinical Epidemiology Program; Institute of Clinical Evaluative Sciences Tanuseputro, Peter; Ottawa Hospital Research Institute, Clinical Epidemiology Program; Institute of Clinical Evaluative Sciences
<b>Primary Subject Heading</b>:	Neurology
Secondary Subject Heading:	Epidemiology, Geriatric medicine, Public health
Keywords:	Dementia < NEUROLOGY, Population projection, Risk stratification, Health behavior

Dementia Population Risk Tool (DemPoRT): Study Protocol for a Predictive Algorithm Assessing Dementia Risk in the Community

Stacey Fisher, MSc[1,2,3] stacey.fisher@uottawa.ca

Amy Hsu, PhD[1,2] ahsu@toh.ca

Nassim Mojaverian, MSc[2] namojaverian@ohri.ca

Monica Taljaard, PhD[1,3]mtaljaard@ohri.ca

Greg Huyer, MSc[1,3] ghuye047@uottawa.ca

Doug Manuel, MD[1,2,3,4,5] dmanuel@ohri.ca

Peter Tanuseputro, MD[1,2,5,6,7] ptanuseputro@ohri.ca

[1] Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

[2] Institute for Clinical Evaluative Sciences, Ottawa, Ontario, Canada

[3] Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada

[4] Statistics Canada, Ottawa, Ontario, Canada

[5] Department of Family Medicine, University of Ottawa, Ottawa, Ontario, Canada

[6] Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada

[7] Bruyère Research Institute, Ottawa, Ontario, Canada

Corresponding author:

Dr. Peter Tanuseputro

1053 Carling Ave.

Box 693
Ottawa, ON K1Y 4E9, Canada
ptanuseputro@ohri.ca
Phone: 613-798-5555

Word Count: 3 723

Keywords: Dementia; population projection; risk stratification; health behavior

ABSTRACT

Introduction: The burden of disease from dementia is a growing global concern as incidence increases exponentially with age and average life expectancy has been increasing around the world. Planning for an aging population requires reliable projections of future dementia prevalence and resource requirements, however, existing population projections are simple and have poor predictive accuracy. The Dementia Population Risk Tool (DemPoRT) will predict incidence of dementia in the population setting using multivariable modeling techniques.

Methods and Analysis: The derivation cohort will consist of elderly Ontario respondents of the Canadian Community Health Survey (CCHS) (2001, 2003, 2005, 2007; 18 764 males and 25 288 females). Pre-specified predictors include sociodemographic, general health, behavioral, functional and health condition variables. Incident dementia will be identified through individual linkage of survey respondents to population-level administrative health care databases (1 797 and 3 281 events, and 117 795 and 166 573 person-years of follow-up, for males and females, respectively until March 31, 2014). Using time of first dementia capture as the primary outcome and death as a competing risk, sex-specific proportional hazards regression models will be estimated. The 2008/2009 CCHS survey will be used for validation (approximately 4 600 males and 6 300 females). Overall calibration and discrimination will be assessed as well as calibration within predefined subgroups of importance to clinicians and policy makers.

Ethics and Dissemination: This study has been approved by the Ottawa Health Science Network Research Ethics Board. DemPoRT results will be submitted for publication in peer-review journals and presented at scientific meetings. The algorithm will be assessable online for both population and individual uses.

Trial Registration Number: ClinicalTrials.gov NCT03155815.

STRENGTHS AND LIMITATIONS

- The Dementia Population Risk Tool (DemPoRT) will be developed and validated using predictors from large population-based community health surveys that are individually linked to routinely-collected health administration data in Ontario. To our knowledge, DemPoRT will be the first population-based algorithm for predicting and projecting dementia incidence.
- DemPoRT will produce improved estimates of future dementia burden, will assess the contribution of specific risk factors to the population risk, and will identify population subgroups at high risk of developing dementia. This information will be used by policymakers to prepare for and reduce dementia impact.
- The analysis plan and predictors have been fully pre-specified to limit the risk of over-fitting and improve the quality of predictions.
- Detailed cognitive testing to ascertain dementia diagnoses is preferable over the use of administrative data, however this is not available or feasible at the population level.
- Although a rigorous approach to model development will be used, further validation will be needed to assess generalizability, and calibration will be required for application in other jurisdictions.

INTRODUCTION

The burden of disease from dementia is a growing global concern as incidence increases exponentially with age and average life expectancy has been increasing around the world[1,2]. Planning for an aging population requires reliable projections of future dementia prevalence and its implications on resource requirements. Existing population projections for dementia, however, are overly simplistic and likely inaccurate[3].

Limitations of Current Dementia Projection Methodology

Almost all existing dementia projections have used extrapolation and macrosimulation methods, which are simplistic and make assumptions that may not hold true into the future[3]. Most extrapolations simply apply current age- and sex-specific prevalence of dementia to future population projections. Macrosimulations use estimates of dementia incidence and mortality, stratified by age and sex, to simulate disease prevalence as the population ages[1,4–6]. Projections from extrapolations incorrectly assume that the risk of mortality among those with and without dementia are equivalent[7,8], and both methods assume that the age and sex-specific prevalence of dementia risk factors will not change with time. The assumption of stable risk factor prevalence is widely thought to be the major source of error in existing dementia projections[3,9–11].

Up to 50% of dementia cases may be attributable to physical inactivity, obesity, diabetes, hypertension, low educational achievement and depression[9,12]. Changing trends of these risk factors over time has the potential to have a strong impact on dementia prevalence. For example, the population prevalence of diabetes and obesity in Canada has been projected to increase, while smoking, hypertension and dyslipidemia have been projected to decline[13]. Consideration of risk factor prevalence is therefore important to improve the accuracy of dementia projections, and simple extrapolations and macrosimulations are often inadequate to incorporate changing risk factors.

Predictive Multivariable Modeling of Dementia Incidence

Another method of dementia projection involves the development of population-based **predictive risk algorithms** that examine the effect of risk factors on dementia incidence. Population-based data that contain detailed exposure information, such as health surveys, are linked at the individual-level to administrative data that capture dementia development. A multivariable model of dementia incidence is derived, validated against external data, and predictive performance is assessed. Counterfactual risk factor levels can be entered in to the algorithm at the population level, or at individual level and summed, to simulate future disease prevalence under different assumptions.

Incorporation of predictive risk algorithms in to microsimulation models such as Statistics Canada's Population Health Models (POHEM) provides additional utility. POHEM dynamically models individual life trajectories of a population representative of Canada including births, deaths and migration, disease incidence and progression, and exposure to risk factors. This facilitates detailed examination of the influence of changing risk factor prevalence on future dementia prevalence and the potential influence of dementia prevention strategies to reduce the population risk. In addition, these algorithms can be used to describe the risk of dementia in the population, assess the contribution of specific risk factors to the population risk, and identify high-risk groups.

The objective of this study is to develop and validate the Dementia Population Risk Tool (DemPoRT) algorithm to predict dementia incidence in the population setting. This will be done using multivariable modelling techniques, linking self-reported risk factors captured by a population-based health survey in Canada with administrative databases across healthcare sectors that capture healthcare diagnosed dementia. To our knowledge, the DemPoRT predictive model will be the first population-based algorithm for predicting and projecting dementia incidence. It will be able to estimate the future burden of dementia using techniques that consider changes in risk factor prevalence and will identify modifiable risk factors that can be targeted by individuals, clinicians and policy makers to reduce dementia incidence more effectively.

METHODS AND ANALYSIS

Study Design

Two DemPoRT models, one for males and females, will be derived and validated using population-based data in Ontario, Canada, a multicultural province with 13.6 million residents. Predictors will be obtained from the Canadian Community Health Surveys (CCHS), and outcomes (i.e., diagnosis of dementia) will be obtained from routinely-collected health care data.

The derivation cohorts will consist of eligible respondents of the 2001, 2003, 2005 and 2007 CCHS (Cycles 1.1, 2.1, 3.1 and 4.1), while validation cohorts will consist of respondents to the 2008/2009 cycle. The CCHS is a national, cross-sectional survey developed by Statistics Canada to collect information related to health and health care utilization of the Canadian population. The survey has a multistage stratified cluster design that represents approximately 98% of the

Canadian population aged 12 years and over and attained an average response rate of 79% over the study period. The CCHS is conducted through telephone and in-person interviews, and all responses are self-reported. The details of survey methodology have been published elsewhere[14]. Survey respondents will be excluded if they are less than 55 years of age at survey administration, self-reported a history of dementia, or are not eligible for Ontario's universal health insurance. If a respondent was included in more than one CCHS cycle, only their earliest survey response will be used.

Outcome

Survey respondents diagnosed with dementia will be identified through individual linkage to several population-based administrative databases at the Institute for Clinical Evaluative Sciences (ICES). Dementia case ascertainment is based on a validated definition: 1 hospital record OR 3 physician claim records at least 30 days apart within a 2-year period OR a dispensing record for a cholinesterase inhibitor from Ontario Drug Benefit (ODB). This definition has a sensitivity of 79.3% and a specificity of 99.1% when validated against emergency medical record (EMR) data[15]. Due to known underdiagnosis of dementia[16,17], we will supplement this definition by adding survey respondents with dementia codes captured on home care and long-term care assessments (dementia flag AND Cognitive Performance Scale [CPS] score ≥2) using the Resident Assessment Instrument-Home Care (RAI-HC) database and the Continuing Care Reporting System (CCRS), respectively. We have found this addition adds substantially (approximately 18%) to the number of dementia cases captured.

Survey respondents with dementia will be excluded if they meet the criteria for dementia within 2 years of survey administration (to remove potentially prevalent cases) or are younger than 65 years of age at the time of dementia diagnosis (to exclude early onset dementia which likely has a different set of risk factors). Eligible survey respondents will be followed from the date of survey administration or age 65, whichever came later, until the earliest date of: dementia ascertainment, death (defined as competing risk), loss to follow-up (defined as loss of healthcare eligibility) or end of study (March 31, 2014).

Sample Size

The male and female derivation cohorts consist of 18 764 and 25 288 respondents, and 117 795 and 166 573 person-years of follow-up, respectively. For predictive models with time to event outcomes the number of participants experiencing the event should exceed 10 times the number of degrees of freedom to ensure adequate sample size[18]. The number of dementia events in the derivation cohort is 1 797 for men and 3 281 for women; therefore, the maximum number of total degrees of freedom for each of the DemPoRT models is 179 and 328, respectively, which we do not anticipate surpassing.

The validation cohorts will consist of approximately 4 600 males and 6 300 females, and 15 000 and 21 000 person-years of follow-up, respectively. Vergouwe *et al*[19] recommend a minimum of 100 events and 100 non-events for external validation studies. We expect approximately 225 events for men and 400 for women in our validation cohort.

Analysis Plan

The analysis plan was developed following guidelines by Harrell[18] and Steyerberg[20] after accessing the derivation data set, but prior to model fitting or descriptive analyses involving exposure-outcome associations. This was done to avoid Type 1 error introduced by data-driven variable selection or model specification. Key considerations of our analysis approach include full pre-specification of the predictor variables, use of flexible functions for continuous predictors, and preserving statistical properties by avoiding data-driven variable selection procedures. Analysis will be conducted using Harrell's Hmisc[21] package of functions in R[22] as well as SAS v9.4.

This study protocol and the reporting of our model estimation results will be guided by the TRIPOD statement for multivariable predictive models[23].

Identification of Predictors

Predictor variables were identified through review of existing predictive algorithms for dementia[9,24–34] and comparison to available data collected in the CCHS. Variable inclusion was informed by consultation with subject-matter experts and the project's advisory team, and informed by our previous work developing predictive models for cardiovascular disease and life expectancy[35,36].

Variables with more than 20% missing values, narrow distributions or insufficient variation were excluded. Obvious cases of redundancy (e.g. alternate definitions of the same underlying behaviour) were not included. A total of 32 predictor variables were identified: 7

sociodemographic, 3 general health, 9 behavioural, 7 functional, 5 health conditions and 1 design variable (CCHS survey cycle). As the effect of dementia risk factors varies by sex, separate models will be derived for men and women. Education, rather than individual income, was selected as a predictor due to several concerns with income including lack of generalizability, measurement error, stability over time and substantial missing values. Indicator variables for smoking status were created to allow the inclusion of smoking pack-years as a continuous predictor. The models will additionally include age interactions with the behavioural, functional and health condition variables as the effect of these risk factors on dementia are expected to vary with age. Detailed definitions and measurement of the predictor variables are presented in Table 1.

Data Cleaning and Coding of Predictors

Continuous variables will be inspected using boxplots and descriptive statistics to determine values outside a plausible range. Values that are clearly erroneous will be corrected, where possible, or set to missing. Continuous predictors with highly skewed distributions will be truncated to the 99.5th percentile. Categorization of continuous variables will be avoided to minimize the loss of predictive information. All data cleaning and coding will occur prior to examining exposure-outcome associations.

Missing Data

As traditional complete cases analyses suffer from inefficiency, selection bias, and other limitations[20], multiple imputation methods will be used to impute missing values using the ‘aregImpute’ function in the HMisc library[21]. This function simultaneously imputes missing values using predictive mean matching and uses bootstrapping to take all aspects of uncertainty in to account. The imputation model will consist of the full list of predictor variables, time to event and censoring variables, as well as auxiliary variables that are not predictors, but may nevertheless be useful in generating imputed values (e.g., income). The final model will be estimated in each of five multiple imputation data sets and the results combined using the rules developed by Rubin and Schenker[37] to account for imputation uncertainty.

Model Specification

Initial sex-specific main effects models will be fit using the pre-specified predictors and an initial degree of freedom allocation for each predictor (Table 1). Decisions on initial degree of freedom allocations were informed by the anticipated importance of each predictor and known dose-response relationships with dementia. Continuous predictors will be flexibly modelled using restricted cubic splines, with the knots placed at fixed quantiles of the distribution (e.g., 5th, 27.5th, 50th, 72.5th, and 95th centiles). Frequency distributions for categorical predictors will be examined and categories with small numbers of respondents will be combined, with analysts blinded to the number of events per category, to avoid instability in the regression analyses. Ordinal variables will be specified as either linear terms or as categorical if the expected association is more complex. Interactions will be restricted to linear terms. The initial model specification, presented in Table 1, includes a total of 86 degrees of freedom (63 main, 23 interaction).

Partial association chi-square statistics for each predictor 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 retain their initial degrees of freedom, while predictors with lower predictive potential will be modeled as simple linear terms or recoded by combining infrequent categories. This process of model specification does not increase the Type I error rate because all predictors will be retained in the full model regardless of their strength of association[18].

Model Estimation

The initial models will be estimated using competing risk Cox proportional hazards regression with time to dementia ascertainment as the outcome and death as a competing risk. Alternative model specifications, including flexible parametric models, will be considered after assessing the validity of model assumptions. All predictors will be centered about their means. A formal check of multicollinearity will be carried out using a variable clustering algorithm[18].

Proportional hazards models assume that the relative risk of the outcome between strata of predictors and the baseline risk must be constant over time. Violation of this assumption has been shown to produce biased results[38] although it has also been argued that the estimated coefficients of time-varying variables can simply be interpreted as an average rather than instantaneous hazard[39]. Plots of raw and smoothed scaled Schoenfeld residuals versus time for each predictor will be assessed to test this assumption and identify non-proportionality. If a violation of this assumption is identified we will consider addition of interaction terms between the predictor and log-transformed time.

Although the risk of overfitting will be minimal due to pre-specification of the models and a large sample size, the need for overfitting adjustment will be assessed. The degree of overfitting will be estimated using the heuristic shrinkage estimator, based on the log likelihood ratio chi-square statistic for the full model[40]. If shrinkage is <0.90, models will be adjusted for overfitting.

Estimation of the Reduced Models

Model pre-specification has advantages in limiting overfitting and spurious statistical significance but can result in a final model that is overly complex, difficult to interpret, and difficult to apply. Unnecessary predictor variables also distort the estimated effects of other predictors making the model more computationally intensive. It is suggested that a more parsimonious model that retains most of the prognostic information and performs as well as or better than the full model can be derived without increasing the Type 1 error rate[18,41]. We will identify a more parsimonious model using a stepdown procedure described by Ambler[41], which involves deleting the variable that results in the smallest decrease in model R^2 until removal leads to an R^2 that is less than a desired level. The reduced model will be evaluated against the full model using Akaike's Information Criterion, and by examining the effect on discrimination and calibration.

DemPoRT will be developed and validated using temporal split samples, however the final regression coefficients will use the full data set to maximize follow-up duration. A cohort-specific intercept and/or interaction term may be included in the final model if the derivation and validation cohorts differ; otherwise, the final combined model will maintain the same predictors and form as the derivation model.

Assessment of Predictive Performance

Predictive performance in the derivation and validation cohorts will be assessed and reported using overall measures of predictive accuracy, discrimination and calibration. Accuracy will be assessed with Nagelkerke's R^2 and the Brier score, and discrimination using the concordance statistic. Model calibration is especially important in the development of prognostic models, as probabilities of future risk are of primary interest[20,42,43]. Calibration will be assessed by comparing the observed and predicted risk of dementia within vigintiles (20 groups of equal frequency) of predicted risk with emphasis on visual inspection of plots rather than formal statistical significance testing, which can be influenced by large sample sizes[19]. Calibration slopes will be generated by regressing the outcome in the validation cohort on the predicted dementia risk, reflecting the combined effect of overfitting to the derivation data as well as true differences in effects of predictors. Deviation of the slope from 1 (perfect calibration) will be tested using a Wald or likelihood ratio test. Calibration within predefined subgroups of importance to clinicians and policy makers (e.g., age group, health behaviour, sociodemographic groups and health conditions) will additionally be evaluated. The clinically relevant standard of calibration was defined as less than 20% difference between observed and predicted estimates within subgroups with a dementia prevalence of at least 5%. All model performance measures will be calculated using the first of the multiply imputed data sets.

Model Presentation

The final regression model, derived from the combined sample of the derivation and validation cohorts, will be presented using estimated hazard ratios and 95% confidence intervals, along with results for the derivation and validation cohorts separately. We have found, however, this usual presentation less meaningful when presenting complex models[35]. To allow interpretation of the estimated effect of each predictor, model behavior will additionally be described using interactive visual tools to display the shape of the effect of each predictor[44]. The regression formula will also be published and used as the basis for web-based implementation.

Analyses Beyond Initial Model Development

We will conduct further analyses exploring the added predictive ability of novel risk factors that were ascertained in single CCHS cycles (e.g. sedentary activity, cognitive stimulation, sleep quality and duration), as well as risk factors that can be ascertained through linkage of additional data sources and similar cohorts (e.g. detailed dietary consumption, lipid levels, blood pressure). In addition, sensitivity analysis of the age at survey administration cutoff used for cohort creation will be performed.

A second, causal model (DemPoRT-C) will also be created to assess the relative contribution of lifestyle, socio-demographic and health factors to dementia incidence. Development will exclude variables believed to be in the causal pathway of dementia occurrence (e.g., self-rated health and functional measures) to reduce the attenuation of hazards from upstream risk factors, but will otherwise be the same as in the predictive model. DemPoRT-C will be applied to the most recent unlinked national CCHS survey.

ETHICS AND MODEL DISSEMINATION

The DemPoRT project advisory committee has been created to ensure that the models meet the needs of knowledge users. This committee has worked with the study team to identify predictors of dementia based on scientific and policy importance and will aid in the identification of important target populations and the establishment of policy-relevant differences for calibration studies.

DemPoRT results will be submitted for publication in peer-review journals and presented at scientific meetings. A web-based individual-level calculator will be created if the models are appropriate for individual use. Although DemPoRT emphasizes risk prediction at the population-level, we have found that individual-level calculators are an effective engagement and translation tool for both the general public and knowledge users.

CONCLUSIONS

To the best of our knowledge, DemPoRT will be the first population-based algorithm for predicting and projecting dementia incidence. The DemPoRT models will produce estimates of future dementia burden that we believe will be more accurate than existing estimates, will assess the contribution of specific risk factors to the population risk, and identify groups at high risk of developing dementia. Although a rigorous approach to model development will be used, further validation will be needed to assess generalizability, and calibration will be required for application in other jurisdictions.

CONTRIBUTIONS

SF drafted and revised the manuscript, and contributed to study design and protocol development. NM contributed to study design, protocol development and provided data/statistical support. AH, MT, DM and GH contributed to the design of the study and protocol development. PT is the lead investigator of the study and was responsible for the conception of the project, the grant application, study design and protocol development. All authors provided critical reviews of the manuscript and approved the final version.

FUNDING

This work is supported by the Canadian Institutes of Health Research operating grant MOP 142237. SF is supported by a doctoral award from the Canadian Institutes of Health Research.

ACKNOWLEDGEMENTS

This project is supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The sponsors have no role in the design or conduct of the study; in the collection, analysis or interpretation of the data; or in the preparation, review or approval or the manuscript. The options, results and conclusions reported are those of the authors and are independent from the funding source. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred.

COMPETING INTERESTS

None declared.

ETHICS APPROVAL

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

REFERNCES

1. Brookmeyer, R., Gray, S. & Kawas, C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. *Am. J. Public Health* **88,** 1337–1342 (1998).
2. Brayne, C. The elephant in the room - healthy brains in later life, epidemiology and public health. *Nat. Rev. Neurosci.* **8,** 233–9 (2007).
3. Norton, S., Matthews, F. E. & Brayne, C. A commentary on studies presenting projections of the future prevalence of dementia. *BMC Public Health* **13,** 1 (2013).
4. Sloane, P. D. *et al.* The public health impact of Alzheimer's disease, 2000-2050: potential implication of treatment advances. *Annu. Rev. Public Health* **23,** 213–31 (2002).
5. Mura, T., Dartigues, J. F. & Berr, C. How many dementia cases in France and Europe? Alternative projections and scenarios 2010-2050. *Eur. J. Neurol.* **17,** 252–259 (2010).
6. Hebert, L. E., Scherr, P. A., Bienias, J. L., Bennett, D. A. & Evans, D. A. Alzheimer disease in the US population: prevalence estimates using the 2000 census. *Arch. Neurol.* **60,** 1119–22 (2003).
7. Brookmeyer, R., Johnson, E., Ziegler-Graham, K. & Arrighi, H. M. Forecasting the global burden of Alzheimer's disease. *Alzheimer's Dement.* **3,** 186–191 (2007).
8. Dewey, M. E. & Chen, C.-M. Neurosis and mortality in persons aged 65 and over living in the community: a systematic review of the literature. *Int. J. Geriatr. Psychiatry* **19,** 554–7 (2004).
9. Norton, S., Matthews, F. E., Barnes, D. E., Yaffe, K. & Brayne, C. Potential for primary prevention of Alzheimer's disease: An analysis of population-based data. *Lancet Neurol.* **13,** 788–794 (2014).

10. Joly, P. *et al.* Prevalence projections of chronic diseases and impact of public health intervention. *Biometrics* **69,** 109–17 (2013).

11. Lee, Y. The recent decline in prevalence of dementia in developed countries: implications for prevention in the Republic of Korea. *J. Korean Med. Sci.* **29,** 913–8 (2014).

12. Barnes, D. E. & Yaffe, K. The projected effect of risk factor reduction on Alzheimer's disease prevalence. *Lancet Neurol.* **10,** 819–828 (2011).

13. Manuel, D. G. *et al.* Projections of preventable risks for cardiovascular disease in Canada to 2021: a microsimulation modelling approach. *C. open* **2,** E94–E101 (2014).

14. Béland, Y. Canadian community health survey--methodological overview. *Heal. reports / Stat. Canada, Can. Cent. Heal. Inf. = Rapp. sur la sant?? / Stat. Canada, Cent. Can. d'information sur la sant??* **13,** 9–14 (2002).

15. Jaakkimainen, R. L. *et al.* Identification of Physician-Diagnosed Alzheimer's Disease and Related Dementias in Population-Based Administrative Data: A Validation Study Using Family Physicians' Electronic Medical Records. *J. Alzheimer's Dis.* **54,** 337–349 (2016).

16. Connolly, A., Gaehl, E., Martin, H., Morris, J. & Purandare, N. Underdiagnosis of dementia in primary care: Variations in the observed prevalence and comparisons to the expected prevalence. *Aging Ment. Health* **15,** 978–984 (2011).

17. Kosteniuk, J. G. *et al.* Incidence and prevalence of dementia in linked administrative health data in Saskatchewan, Canada: a retrospective cohort study. *BMC Geriatr.* **15,** 73 (2015).

18. Harrell, F. E. *Regression Modeling Strategies with applicaitons to linear models, logistic regression and survival analysis*. (Springer, 2001).

19. Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. C. & Habbema, J. D. F. Substantial

effective sample sizes were required for external validation studies of predictive logistic regression models. *J. Clin. Epidemiol.* **58,** 475–483 (2005).

20. Steyerberg, E. W. *Clinical Prediction Models*. (Springer, 2009).

21. Harrell, F. E. *Hmisc: Harrell Miscellaneous. R package version 4.0-2*. (2016).

22. R Core Team. R: A language and environment for statistical computing. (2016).

23. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur. Urol.* **67,** 1142–1151 (2015).

24. Walters, K. *et al.* Predicting dementia risk in primary care: development and validation of the Dementia Risk Score using routinely collected data. *BMC Med.* **14,** 1–12 (2016).

25. Anstey, K. J. *et al.* A self-report risk index to predict occurrence of dementia in three independent cohorts of older adults: The ANU-ADRI. *PLoS One* **9,** (2014).

26. Exalto, L. G. *et al.* Midlife risk score for the prediction of dementia four decades later. *Alzheimer's Dement.* **10,** 562–570 (2014).

27. Jessen, F. *et al.* Prediction of dementia in primary care patients. *PLoS One* **6,** e16852 (2011).

28. Reitz, C. *et al.* A summary risk score for the prediction of Alzheimer disease in elderly persons. *Arch. Neurol.* **67,** 835–41 (2010).

29. Barnes, D. E. *et al.* Predicting risk of dementia in older adults: The late-life dementia risk index. *Neurology* **73,** 173–179 (2009).

30. Kivipelto, M. *et al.* Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study. *Lancet Neurol.* **5,** 735–741 (2006).

31. Song, X., Mitnitski, A. & Rockwood, K. Nontraditional risk factors combine to predict Alzheimer disease and dementia. *Neurology* **77,** 227–234 (2011).

32. Tierney, M. C., Moineddin, R. & McDowell, I. Prediction of all-cause dementia using neuropsychological tests within 10 and 5 years of diagnosis in a community-based sample. *J. Alzheimer's Dis.* **22,** 1231–1240 (2010).

33. Exalto, L. G. *et al.* Risk score for prediction of 10 year dementia risk in individuals with type 2 diabetes: A cohort study. *Lancet Diabetes Endocrinol.* **1,** 183–190 (2013).

34. Chary, E. *et al.* Short-versus long-term prediction of dementia among subjects with low and high educational levels. *Alzheimer's Dement.* **9,** 562–571 (2013).

35. Taljaard, M. *et al.* Cardiovascular Disease Population Risk Tool (CVDPoRT): predictive algorithm for assessing CVD risk in the community setting. A study protocol. *BMJ Open* **4,** e006701 (2014).

36. Manuel, D. G. *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.* **13,** 1–27 (2016).

37. Rubin, D. B. & Schenker, N. Multiple imputation in health-care databases: an overview and some applications. *Stat. Med.* **10,** 585–98 (1991).

38. Therneau, T. M., Grambsch, P. M. & Fleming, T. R. Martingale-based residuals for survival models. *Biometrika* **77,** 147–160 (1990).

39. Allison, P. D. *Survival analysis using SAS : A practical guide*. (SAS Institute, 2010).

40. Van Houwelingen, J. C. & Le, C. S. Predictive value of statistical models. *Stat Med* **9,** 1303–1325 (1990).

41. Ambler, G., Brady, A. R. & Royston, P. Simplifying a prognostic model: A simulation

study based on clinical data. *Stat. Med.* **21,** 3803–3822 (2002).

42. Cook, N. R. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin. Chem.* **54,** 17–23 (2008).

43. Cook, N. R. Comment: Measures to summarize and compare the predictive capacity of markers. *Int. J. Biostat.* **6,** Article 22; discussion Article 25 (2010).

44. Krause, J., Perer, A. & Bertini, E. Using Visual Analytics to Interpret Predictive Machine Learning Models. *arXiv Prepr. arXiv1606.05685.* (2016).

Table 1. Pre-specification of predictor variables for DemPoRT with initial degrees of freedom (df) allocation

Variable	Scale	Initial Variable Specification	df
Socio-demographic Factors			
Age	Continuous	**5 knot spline:** Valid range: 55-102 (male), 55-101 (female)	4
Sex	Categorical	**Stratified:** Male; Female	NA
Ethnicity	Categorical	**7 categories:** 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
Education	Categorical	**4 categories:** Less than secondary school; Secondary school graduation; Some postsecondary; Postsecondary graduation	3
Marital Status	Categorical	**4 categories:** Now married/Common-law; Separated/Divorced; Widowed; Single	3
Neighborhood Social and Material Deprivation (Pampalon et al. 2009)	Ordinal	**3 categories:** Low (1st or 2nd quintile); High 4th or 5th quintile; Moderate (3rd quintile)	2
General Health			
Sense of belonging to local community	Ordinal	**4 categories:** Very strong; Somewhat strong; Somewhat weak; Very weak	3
Self-perceived stress	Ordinal	**5 categories:** Not at all stressful; Not very stressful; A bit stressful; Quite a bit stressful; Extremely stressful	4
Self-rated health	Ordinal	**5 categories:** Poor; Fair; Good; Very Good; Excellent	4
Health Behaviors			
Pack years of smoking	Continuous	**3 knot spline:** Valid range: 0-112 (male), 0-78 (female)	2
Smoking status	Categorical	**4 categories:** 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	**3 knot spline:** Valid range: 0-50 (male), 0-24 (female)	2
Former drinker	Dichotomous	Yes; No	1
Consumption of fruit, salad, carrot and other vegetables (average daily frequency)	Continuous	**3 knot spline:** Valid range: 0-48 (male), 0-31 (female)	2
Potato consumption (average daily frequency)	Continuous	**3 knot spline:** Valid range: 0-2	2
Juice consumption (average daily consumption	Continuous	**3 knot spline:** Valid range: 0-6 (male), 0-5 (female)	2
Leisure physical activity (average daily METs (kcal/kg/day))	Continuous	**3 knot spline:** Valid range: 0-16 (male), 0-12 (female)	2
Functional Measures			
Personal hygiene and care	Dichotomous	Does not need help; Needs help	1
Locomotion in the home	Dichotomous	Does not need help; Needs help	1
Meal preparation	Dichotomous	Does not need help; Needs help	1
Running errands	Dichotomous	Does not need help; Needs help	1
Ordinary housework	Dichotomous	Does not need help; Needs help	1
Heavy housework	Dichotomous	Does not need help; Needs help	1
Finances	Dichotomous	Does not need help; Needs help	1
Health Conditions			
Heart disease	Dichotomous	Yes; No	1
Stroke	Dichotomous	Yes; No	1
Diabetes	Dichotomous	Yes; No	1
Mood disorder	Dichotomous	Yes; No	1
High blood pressure	Dichotomous	Yes; No	1
Body mass index	Continuous	**3 knot spline:** Valid range: 10-44 (male), 10-47 (female)	2
Design			
Survey year	Ordinal	**4 categories:** 2000/01, 2002/03, 2004/05, 2006/07	3

DemPoRT, Dementia Population Risk Tool; df, degrees of freedom; MET, metabolic equivalent task

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	3
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	5,6
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	7
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	7,8
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	7,8
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	7,8
	5b	D;V	Describe eligibility criteria for participants.	7-9
	5c	D;V	Give details of treatments received, if relevant.	NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	8,9
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	10
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	10,11, 25
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	10
Sample size	8	D;V	Explain how the study size was arrived at.	7-9
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	11,12
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	12,13, 25
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	13,14
	10c	V	For validation, describe how the predictions were calculated.	15
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	15
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	15
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	15,16
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	9
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	NA
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	NA
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	NA
	15b	D	Explain how to the use the prediction model.	NA
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	NA
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	NA
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	4,17
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	NA
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	NA
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	5-7
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	4,17
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	18

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

BMJ Open

Dementia Population Risk Tool (DemPoRT): Study Protocol for a Predictive Algorithm Assessing Dementia Risk in the Community

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018018.R1
Article Type:	Protocol
Date Submitted by the Author:	01-Aug-2017
Complete List of Authors:	Fisher, Stacey; Ottawa Hospital Research Institute; University of Ottawa, Department of Epidemiology and Community Medicine Hsu, Amy; Ottawa Hospital Research Institute; Institute of Clinical Evaluative Sciences Mojaverian, Nassim; Institute of Clinical Evaluative Sciences Taljaard, Monica; Ottawa Hospital Research Institute, Clinical Epidemiology Program; University of Ottawa, Department of Epidemiology and Community Medicine Huyer, Greg; University of Ottawa Telfer School of Management Manuel, Douglas; Ottawa Hospital Research Institute, Clinical Epidemiology Program; Institute of Clinical Evaluative Sciences Tanuseputro, Peter; Ottawa Hospital Research Institute, Clinical Epidemiology Program; Institute of Clinical Evaluative Sciences
<b>Primary Subject Heading</b>:	Neurology
Secondary Subject Heading:	Epidemiology, Geriatric medicine, Public health
Keywords:	Dementia < NEUROLOGY, Population projection, Risk stratification, Health behavior

Dementia Population Risk Tool (DemPoRT): Study Protocol for a Predictive Algorithm Assessing Dementia Risk in the Community

Stacey Fisher, MSc[1,2,3] stacey.fisher@uottawa.ca

Amy Hsu, PhD[1,2] ahsu@toh.ca

Nassim Mojaverian, MSc[2] namojaverian@ohri.ca

Monica Taljaard, PhD[1,3]mtaljaard@ohri.ca

Gregory Huyer, PhD[4] ghuye047@uottawa.ca

Doug Manuel, MD[1,2,3,5,6] dmanuel@ohri.ca

Peter Tanuseputro, MD[1,2,6,7,8] ptanuseputro@ohri.ca

[1] Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

[2] Institute for Clinical Evaluative Sciences, Ottawa, Ontario, Canada

[3] Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada

[4] Telfer School of Management, University of Ottawa, Ottawa, Ontario, Canada

[5] Statistics Canada, Ottawa, Ontario, Canada

[6] Department of Family Medicine, University of Ottawa, Ottawa, Ontario, Canada

[7] Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada

[8] Bruyère Research Institute, Ottawa, Ontario, Canada

Corresponding author:

Dr. Peter Tanuseputro

1053 Carling Ave.
Box 693
Ottawa, ON K1Y 4E9, Canada
ptanuseputro@ohri.ca
Phone: 613-798-5555

Word Count: 4 046

Keywords: Dementia; population projection; risk stratification; health behavior

ABSTRACT

Introduction: The burden of disease from dementia is a growing global concern as incidence increases dramatically with age and average life expectancy has been increasing around the world. Planning for an aging population requires reliable projections of dementia prevalence; however, existing population projections are simple and have poor predictive accuracy. The Dementia Population Risk Tool (DemPoRT) will predict incidence of dementia in the population setting using multivariable modeling techniques, and will be used to project dementia prevalence.

Methods and Analysis: The derivation cohort will consist of elderly Ontario respondents of the Canadian Community Health Survey (CCHS) (2001, 2003, 2005, 2007; 18 764 males and 25 288 females). Pre-specified predictors include sociodemographic, general health, behavioral, functional and health condition variables. Incident dementia will be identified through individual linkage of survey respondents to population-level administrative health care databases (1 797 and 3 281 events, and 117 795 and 166 573 person-years of follow-up, for males and females, respectively until March 31, 2014). Using time of first dementia capture as the primary outcome and death as a competing risk, sex-specific proportional hazards regression models will be estimated. The 2008/2009 CCHS survey will be used for validation (approximately 4 600 males and 6 300 females). Overall calibration and discrimination will be assessed as well as calibration within predefined subgroups of importance to clinicians and policy makers.

Ethics and Dissemination: Research ethics approval has been granted by the Ottawa Health Science Network Research Ethics Board. DemPoRT results will be submitted for publication in peer-review journals and presented at scientific meetings. The algorithm will be assessable online for both population and individual uses.

Trial Registration Number: ClinicalTrials.gov NCT03155815.

STRENGTHS AND LIMITATIONS

- The Dementia Population Risk Tool (DemPoRT) will be developed and validated using predictors from large population-based community health surveys that are individually linked to routinely-collected health administration data in Ontario. To our knowledge, DemPoRT will be the first algorithm designed to predict and project dementia incidence at the population-level.
- DemPoRT will be used to produce improved estimates of future dementia burden, will assess the contribution of specific risk factors to the population risk, and will identify population subgroups at high risk of developing dementia. This information will be used by policymakers to prepare for and reduce dementia impact.
- The analysis plan and predictors have been fully pre-specified to limit the risk of over-fitting and improve the quality of predictions.
- Detailed cognitive testing to ascertain dementia diagnoses is preferable over the use of administrative data, however this is not available or feasible at the population level.
- Although a rigorous approach to model development will be used, further validation will be needed to assess generalizability, and calibration will be required for application in other jurisdictions.

INTRODUCTION

The burden of disease from dementia is a growing global concern as incidence increases dramatically with age and average life expectancy has been increasing around the world[1,2]. Planning for an aging population requires reliable projections of dementia burden and the implications for resource requirements. Existing population-level projections for dementia, however, are overly simplistic and likely inaccurate[3].

Limitations of Current Dementia Projection Methodology

Almost all existing dementia projections have used extrapolation and macrosimulation methods, which are simplistic and make assumptions that may not hold true into the future[3]. Most extrapolations simply apply current age- and sex-specific prevalence estimates of dementia to future population projections. Macrosimulations typically use estimates of dementia incidence and mortality, stratified by age and sex, to simulate disease prevalence as the population ages[1,4–6]. Projections from extrapolations incorrectly assume that the risk of mortality among those with and without dementia are equivalent[7,8], and both methods assume that the age and sex-specific prevalence of dementia risk factors will not change with time. The assumption of stable risk factor prevalence is widely thought to be the major source of error in existing dementia projections[3,9–11].

Changing trends of dementia risk factors has the potential to have a dramatic impact on dementia prevalence estimates, as up to 50% of dementia cases have been attributed to modifiable factors[9,12], and the prevalence of several factors has been projected to change significantly in the near future. For example, the population prevalence of diabetes and obesity in Canada has been projected to increase, while smoking, hypertension and dyslipidemia have been projected to decline[13]. Consideration of risk factor prevalence is therefore important to improve the accuracy of dementia projections, and simple extrapolations and macrosimulations are often inadequate.

Predictive Multivariable Modeling of Dementia Incidence

Population-based **predictive risk algorithms** examine the effect of risk factors on dementia incidence, and can be used for dementia burden projection. Population-based data that contain detailed risk factor information, such as health surveys, are linked at the individual-level to administrative data that capture dementia development. A multivariable model of dementia incidence is derived, validated against external data, and predictive performance is assessed. Once developed, the algorithm can be used to project disease incidence and prevalence. To obtain prevalence projections, the algorithm can be integrated in to a microsimulation model such as Statistics Canada's Population Health Models (POHEM). POHEM dynamically models individual life trajectories of a population representative of Canada including births, deaths and migration, disease incidence and progression, and exposure to risk factors, facilitating detailed examination of the influence of changing risk factor prevalence on future dementia prevalence.

Predictive risk algorithms can also be used to describe the risk of dementia in the population, assess the contribution of specific risk factors to the population risk, identify high-risk groups, and evaluate risk reduction strategies.

Existing Dementia Prediction Models

Many models have been developed to predict risk of dementia[14–26], most with the primary goal of identifying individuals in the clinical setting at high risk. They have varying discriminative ability (c-statistics ranging from 0.49[16] to 0.89[17]) and have generally been derived from small samples, rarely including more than a few thousand individuals. Existing models are therefore simplistic, including few predictors and rarely including interaction or non-linear terms Existing models thus facilitates understanding and use by physicians in clinical practice, but limits discriminatory ability and predictive accuracy. Walters et al[26] developed an algorithm for predicting 5-year dementia risk among individuals 60-79 years of age in the United Kingdom using an enormous derivation dataset of 800 000 individuals, and a simple model. The derivation model had a c-statistic of 0.84 (95% CI: 0.81, 0.87), but a low positive predictive value at most risk thresholds, and therefore is poor at identifying those at high risk of dementia. Additionally, as most dementia risk models are intended for use in the clinical setting, many include results from neuropsycological tests[17–23], MRI findings[18] and APOE genotype[18,24,25]. The inclusion of these variables, however, limits the application of these models as these variables are not available at the population-level.

The objective of this study is to develop and validate the Dementia Population Risk Tool (DemPoRT) algorithm to predict dementia incidence in the population setting. This will be done using multivariable modelling techniques, linking self-reported risk factors captured by a population-based health survey in Canada with administrative databases across healthcare sectors that capture healthcare diagnosed dementia. DemPoRT will be developed with a using a large population-based dataset using only variables that are available at the population-level, allowing for population-level application. DemPoRT will also utilize many methodological improvements over existing models. This protocol pre-specifies the predictor variables and analytic plan for model development, reducing the potential for overfitting and bias, and improving transparency. Interaction terms and flexible functions for continuous predictors will be investigated, increasing potential discriminative ability. The pre-specified analytic plan avoids data-driven variable selection procedures, further reducing the potential for bias.

To our knowledge, the DemPoRT predictive model will be the first algorithm designed to predict and project dementia incidence at the population-level. It will be used to estimate the future burden of dementia using techniques that consider changes in risk factor prevalence and will identify modifiable risk factors that can be targeted by individuals, clinicians and policy makers to reduce the burden of dementia.

METHODS AND ANALYSIS

Study Design

Two DemPoRT models, one for males and females, will be derived and validated using population-based data in Ontario, Canada, a multicultural province with 13.6 million residents. Predictors will be obtained from the Canadian Community Health Surveys (CCHS), and outcomes (i.e., diagnosis of dementia) will be obtained from routinely-collected health care data.

The derivation cohorts will consist of eligible respondents of the 2001, 2003, 2005 and 2007 CCHS (Cycles 1.1, 2.1, 3.1 and 4.1), while validation cohorts will consist of respondents to the 2008/2009 cycle. The CCHS is a national, cross-sectional survey developed by Statistics Canada to collect information related to health and health care utilization of the Canadian population.

The survey has a multistage stratified cluster design that represents approximately 98% of the Canadian population aged 12 years and over and attained an average response rate of 79% over the study period. The CCHS is conducted through telephone and in-person interviews, and all responses are self-reported. The details of survey methodology have been published elsewhere[27]. Survey respondents will be excluded if they are less than 55 years of age at survey administration, self-reported a history of dementia, or are not eligible for Ontario's universal health insurance. If a respondent was included in more than one CCHS cycle, only their earliest survey response will be used.

Outcome

Survey respondents diagnosed with dementia will be identified through individual linkage to several population-based administrative databases at the Institute for Clinical Evaluative Sciences (ICES). Dementia case ascertainment is based on a validated definition: 1 hospital record OR 3 physician claim records at least 30 days apart within a 2-year period OR a dispensing record for a cholinesterase inhibitor from Ontario Drug Benefit (ODB). This definition has a sensitivity of 79.3% and a specificity of 99.1% when validated against emergency medical record (EMR) data[28]. Due to known underdiagnosis of dementia[29,30], we will supplement this definition by adding survey respondents with dementia codes captured on home care and long-term care assessments (dementia flag AND Cognitive Performance Scale [CPS] score $\geq$2) using the Resident Assessment Instrument-Home Care (RAI-HC) database and the Continuing Care Reporting System (CCRS), respectively. We have found this addition adds substantially (approximately 18%) to the number of dementia cases captured.

Survey respondents with dementia will be excluded if they meet the criteria for dementia within 2 years of survey administration (to remove potentially prevalent cases) or are younger than 65 years of age at the time of dementia diagnosis (to exclude early onset dementia which likely has a different set of risk factors). Eligible survey respondents will be followed from the date of survey administration or age 65, whichever came later, until the earliest date of: dementia ascertainment, death (defined as competing risk), loss to follow-up (defined as loss of healthcare eligibility) or end of study (March 31, 2014).

Sample Size

The male and female derivation cohorts consist of 18 764 and 25 288 respondents, and 117 795 and 166 573 person-years of follow-up, respectively. For predictive models with time to event outcomes the number of participants experiencing the event should exceed 10 times the number of degrees of freedom to ensure adequate sample size[31]. The number of dementia events in the derivation cohort is 1 797 for men and 3 281 for women; therefore, the maximum number of total degrees of freedom for each of the DemPoRT models is 179 and 328, respectively, which we do not anticipate surpassing.

The validation cohorts will consist of approximately 4 600 males and 6 300 females, and 15 000 and 21 000 person-years of follow-up, respectively. Vergouwe *et al*[32] recommend a minimum of 100 events and 100 non-events for external validation studies. We expect approximately 225 events for men and 400 for women in our validation cohort.

Analysis Plan

The analysis plan was developed following guidelines by Harrell[31] and Steyerberg[33] after accessing the derivation data set, but prior to model fitting or descriptive analyses involving exposure-outcome associations. This was done to avoid Type 1 error introduced by data-driven variable selection or model specification. Key considerations of our analysis approach include full pre-specification of the predictor variables, use of flexible functions for continuous predictors, and preserving statistical properties by avoiding data-driven variable selection procedures. Analysis will be conducted using Harrell's Hmisc[34] package of functions in R[35] as well as SAS v9.4.

This study protocol and the reporting of our model estimation results will be guided by the TRIPOD statement for multivariable predictive models[36].

Identification of Predictors

Predictor variables were identified through review of existing predictive algorithms for dementia[9,16,18–22,24–26,37,38] and comparison to available data collected in the CCHS. Variable inclusion was informed by consultation with subject-matter experts and the project's advisory team, and informed by our previous work developing predictive models for cardiovascular disease and life expectancy[39,40].

Variables with narrow distributions or insufficient variation were excluded. Obvious cases of redundancy (e.g. alternate definitions of the same underlying behavior) were not included. A total of 32 predictor variables were identified: 7 sociodemographic, 3 general health, 9 behavioral, 7 functional, 5 health conditions and 1 design variable (CCHS survey cycle). As the effect of dementia risk factors varies by sex, separate models will be derived for men and women. Education, rather than individual income, was selected as a predictor due to several concerns with income including lack of generalizability, measurement error, stability over time and substantial missing values. Neighborhood social and material deprivation is captured using Pampalon's deprivation index[41]. Indicator variables for smoking status were created to allow the inclusion of smoking pack-years as a continuous predictor. The models will additionally include age interactions with the behavioral, functional and health condition variables as the effect of these risk factors on dementia are expected to vary with age. Detailed definitions and measurement of the predictor variables are presented in Table 1.

Data Cleaning and Coding of Predictors

Continuous variables will be inspected using boxplots and descriptive statistics to determine values outside a plausible range. Values that are clearly erroneous will be corrected, where possible, or set to missing. Continuous predictors with highly skewed distributions will be truncated to the 99.5th percentile. Categorization of continuous variables will be avoided to minimize the loss of predictive information. All data cleaning and coding will occur prior to examining exposure-outcome associations.

Missing Data

As traditional complete cases analyses suffer from inefficiency, selection bias, and other limitations[33], multiple imputation methods will be used to impute missing values using the 'aregImpute' function in the HMisc library[34]. This function simultaneously imputes missing values using predictive mean matching and uses bootstrapping to take all aspects of uncertainty in to account. The imputation model will consist of the full list of predictor variables, time to event and censoring variables, as well as auxiliary variables that are not predictors, but may nevertheless be useful in generating imputed values (e.g., income). The final model will be estimated in each of five multiple imputation data sets and the results combined using the rules developed by Rubin and Schenker[42] to account for imputation uncertainty.

Model Specification

Initial sex-specific main effects models will be fit using the pre-specified predictors and an initial degree of freedom allocation for each predictor (Table 1). Decisions on initial degree of freedom allocations were informed by the anticipated importance of each predictor and known dose-response relationships with dementia. Continuous predictors will be flexibly modelled using restricted cubic splines, with the knots placed at fixed quantiles of the distribution (e.g., 5th, 27.5th, 50th, 72.5th, and 95th centiles). Frequency distributions for categorical predictors will be examined and categories with small numbers of respondents will be combined, with analysts blinded to the number of events per category, to avoid instability in the regression analyses. Ordinal variables will be specified as either linear terms or as categorical if the expected association is more complex. Interactions will be restricted to linear terms. The initial model specification, presented in Table 1, includes a total of 86 degrees of freedom (63 main, 23 interaction).

Partial association chi-square statistics for each predictor 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 retain their initial degrees of freedom, while predictors with lower predictive potential will be modeled as simple linear terms or recoded by combining infrequent categories. This process of model specification does not increase the Type I error rate because all predictors will be retained in the full model regardless of their strength of association[31].

Model Estimation

The initial models will be estimated using competing risk Cox proportional hazards regression with time to dementia ascertainment as the outcome and death as a competing risk. Alternative model specifications, including subdistribution hazard and flexible parametric models, will be considered. All predictors will be centered about their means. A formal check of multicollinearity will be carried out using a variable clustering algorithm[31].

Proportional hazards models assume that the relative risk of the outcome between strata of predictors and the baseline risk must be constant over time. Violation of this assumption has been shown to produce biased results[43] although it has also been argued that the estimated coefficients of time-varying variables can simply be interpreted as an average rather than instantaneous hazard[44]. Plots of raw and smoothed scaled Schoenfeld residuals versus time for each predictor will be assessed to test this assumption and identify non-proportionality. If a violation of this assumption is identified we will consider addition of interaction terms between the predictor and log-transformed time.

Although the risk of overfitting will be minimal due to pre-specification of the models and a large sample size, the need for overfitting adjustment will be assessed. The degree of overfitting will be estimated using the heuristic shrinkage estimator, based on the log likelihood ratio chi-square statistic for the full model[45]. If shrinkage is <0.90, models will be adjusted for overfitting.

Estimation of the Reduced Models

Model pre-specification has advantages in limiting overfitting and spurious statistical significance but can result in a final model that is overly complex, difficult to interpret, and difficult to apply. Unnecessary predictor variables also distort the estimated effects of other predictors making the model more computationally intensive. It is suggested that a more parsimonious model that retains most of the prognostic information and performs as well as or better than the full model can be derived without increasing the Type 1 error rate[31,46]. We will identify a more parsimonious model using a stepdown procedure described by Ambler[46], which involves deleting the variable that results in the smallest decrease in model R^2 until removal leads to an R^2 that is less than a desired level. The reduced model will be evaluated against the full model using Akaike's Information Criterion, and by examining the effect on discrimination and calibration.

DemPoRT will be developed and validated using temporal split samples, however the final regression coefficients will use the full data set to maximize follow-up duration. A cohort-specific intercept and/or interaction term may be included in the final model if the derivation and validation cohorts differ; otherwise, the final combined model will maintain the same predictors and form as the derivation model.

Assessment of Predictive Performance

Predictive performance in the derivation and validation cohorts will be assessed and reported using overall measures of predictive accuracy, discrimination and calibration. Accuracy will be assessed with Nagelkerke's R^2 [47] and the Brier score[48]. Discrimination will be assessed using the concordance statistic. Model calibration is especially important in the development of prognostic models, as probabilities of future risk are of primary interest[33,49,50]. Calibration will be assessed by comparing the observed and predicted risk of dementia within vigintiles (20 groups of equal frequency) of predicted risk with emphasis on visual inspection of plots rather than formal statistical significance testing, which can be influenced by large sample sizes[32]. Calibration slopes will be generated by regressing the outcome in the validation cohort on the predicted dementia risk, reflecting the combined effect of overfitting to the derivation data as well as true differences in effects of predictors. Deviation of the slope from 1 (perfect calibration) will be tested using a Wald or likelihood ratio test. Calibration within predefined subgroups of importance to clinicians and policy makers (e.g., age group, health behavior, sociodemographic groups and health conditions) will additionally be evaluated. The clinically relevant standard of calibration was defined as less than 20% difference between observed and predicted estimates within subgroups with a dementia prevalence of at least 5%. All model performance measures will be calculated using the first of the multiply imputed data sets.

Model Presentation

The final regression model, derived from the combined sample of the derivation and validation cohorts, will be presented using estimated hazard ratios and 95% confidence intervals, along with results for the derivation and validation cohorts separately. We have found, however, this usual presentation less meaningful when presenting complex models[39]. To allow interpretation of the estimated effect of each predictor, model behavior will additionally be described using interactive visual tools to display the shape of the effect of each predictor[51]. The regression formula will also be published and used as the basis for web-based implementation.

Analyses Beyond Initial Model Development

We will conduct further analyses exploring the added predictive ability of novel risk factors that were ascertained in single CCHS cycles (e.g. sedentary activity, cognitive stimulation, sleep quality and duration, deafness), as well as risk factors that can be ascertained through linkage of additional data sources and similar cohorts (e.g. air pollution, detailed dietary consumption, lipid levels, blood pressure). In addition, sensitivity analysis of the age at survey administration cutoff used for cohort creation will be performed.

Once developed, DemPoRT will be used to project dementia incidence under different assumptions by entering counterfactual risk factor levels in to the algorithm at the population level, or at individual level and summed, and will be integrated in to POHEM for microsimulation modelling of prevalence projections.

A second, causal model (DemPoRT-C) will also be created to assess the relative contribution of lifestyle, socio-demographic and health factors to dementia incidence. Development will exclude variables believed to be in the causal pathway of dementia occurrence (e.g., self-rated health and functional measures) to reduce the attenuation of hazards from upstream risk factors, but will otherwise be the same as in the predictive model. DemPoRT-C will be applied to the most recent unlinked national CCHS survey.

ETHICS AND MODEL DISSEMINATION

The DemPoRT project advisory committee has been created to ensure that the models meet the needs of knowledge users. This committee has worked with the study team to identify predictors of dementia based on scientific and policy importance and will aid in the identification of important target populations and the establishment of policy-relevant differences for calibration studies.

DemPoRT results will be submitted for publication in peer-review journals and presented at scientific meetings. A web-based individual-level calculator will be created if the models are appropriate for individual use. Although DemPoRT emphasizes risk prediction at the population-level, we have found that individual-level calculators are an effective engagement and translation tool for both the general public and knowledge users.

CONCLUSIONS

To the best of our knowledge, DemPoRT will be the first population-based algorithm designed to predicting and projecting dementia incidence at the population level. The DemPoRT models will produce estimates of future dementia burden that we believe will be more accurate than existing estimates, will assess the contribution of specific risk factors to the population risk, and identify groups at high risk of developing dementia. Although a rigorous approach to model development will be used, further validation will be needed to assess generalizability, and calibration will be required for application in other jurisdictions.

CONTRIBUTIONS

SF drafted and revised the manuscript, and contributed to study design and protocol development. NM contributed to study design, protocol development and provided data/statistical support. AH, MT, DM and GH contributed to the design of the study and protocol development. PT is the lead investigator of the study and was responsible for the conception of the project, the grant application, study design and protocol development. All authors provided critical reviews of the manuscript and approved the final version.

FUNDING

This work is supported by the Canadian Institutes of Health Research operating grant MOP 142237. SF is supported by a doctoral award from the Canadian Institutes of Health Research.

ACKNOWLEDGEMENTS

This project is supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The sponsors have no role in the design or conduct of the study; in the collection, analysis or interpretation of the data; or in the preparation, review or approval or the manuscript. The options, results and conclusions reported are those of the authors and are independent from the funding source. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred.

COMPETING INTERESTS

None declared.

ETHICS APPROVAL

Research ethics approval has been granted by the Ottawa Health Science Network Research Ethics Board.

REFERNCES

1. Brookmeyer, R., Gray, S. & Kawas, C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. *Am. J. Public Health* **88,** 1337–1342 (1998).

2. Brayne, C. The elephant in the room - healthy brains in later life, epidemiology and public health. *Nat. Rev. Neurosci.* **8,** 233–9 (2007).

3. Norton, S., Matthews, F. E. & Brayne, C. A commentary on studies presenting projections of the future prevalence of dementia. *BMC Public Health* **13,** 1 (2013).

4. Sloane, P. D. *et al.* The public health impact of Alzheimer's disease, 2000-2050: potential implication of treatment advances. *Annu. Rev. Public Health* **23,** 213–31 (2002).

5. Mura, T., Dartigues, J. F. & Berr, C. How many dementia cases in France and Europe? Alternative projections and scenarios 2010-2050. *Eur. J. Neurol.* **17,** 252–259 (2010).

6. Hebert, L. E., Scherr, P. A., Bienias, J. L., Bennett, D. A. & Evans, D. A. Alzheimer disease in the US population: prevalence estimates using the 2000 census. *Arch. Neurol.* **60,** 1119–22 (2003).

7. Brookmeyer, R., Johnson, E., Ziegler-Graham, K. & Arrighi, H. M. Forecasting the global burden of Alzheimer's disease. *Alzheimer's Dement.* **3,** 186–191 (2007).

8. Dewey, M. E. & Chen, C.-M. Neurosis and mortality in persons aged 65 and over living in the community: a systematic review of the literature. *Int. J. Geriatr. Psychiatry* **19,** 554–7 (2004).

9. Norton, S., Matthews, F. E., Barnes, D. E., Yaffe, K. & Brayne, C. Potential for primary prevention of Alzheimer's disease: An analysis of population-based data. *Lancet Neurol.* **13,** 788–794 (2014).

10. Joly, P. *et al.* Prevalence projections of chronic diseases and impact of public health intervention. *Biometrics* **69,** 109–17 (2013).

11. Lee, Y. The recent decline in prevalence of dementia in developed countries: implications for prevention in the Republic of Korea. *J. Korean Med. Sci.* **29,** 913–8 (2014).

12. Barnes, D. E. & Yaffe, K. The projected effect of risk factor reduction on Alzheimer's disease prevalence. *Lancet. Neurol.* **10,** 819–28 (2011).

13. Manuel, D. G. *et al.* Projections of preventable risks for cardiovascular disease in Canada to 2021: a microsimulation modelling approach. *C. Open* **2,** E94–E101 (2014).

14. Tang, E. Y. H. *et al.* Current developments in dementia risk prediction modelling: An updated systematic review. *PLoS One* **10,** 1–31 (2015).

15. Stephan, B. C. M., Kurth, T., Matthews, F. E., Brayne, C. & Dufouil, C. Dementia risk prediction in the population: are screening models accurate? *Nat. Rev. Neurol.* **6,** 318–326 (2010).

16. Anstey, K. J. *et al.* A self-report risk index to predict occurrence of dementia in three independent cohorts of older adults: The ANU-ADRI. *PLoS One* **9,** (2014).

17. Wolfsgruber, S. *et al.* The CERAD neuropsychological assessment battery total score detects and predicts alzheimer disease dementia with high diagnostic accuracy. *Am. J. Geriatr. Psychiatry* **22,** 1017–1028 (2014).

18. Barnes, D. E. *et al.* Predicting risk of dementia in older adults: The late-life dementia risk index. *Neurology* **73,** 173–179 (2009).

19. Chary, E. *et al.* Short- versus long-term prediction of dementia among subjects with low and high educational levels. *Alzheimers. Dement.* **9,** 562–71 (2013).

20. Jessen, F. *et al.* Prediction of Dementia in Primary Care Patients. *PLoS One* **6,** e16852

(2011).

21. Song, X., Mitnitski, A. & Rockwood, K. Nontraditional risk factors combine to predict Alzheimer disease and dementia. *Neurology* **77,** 227–234 (2011).

22. Tierney, M. C., Moineddin, R. & McDowell, I. Prediction of all-cause dementia using neuropsychological tests within 10 and 5 years of diagnosis in a community-based sample. *J. Alzheimer's Dis.* **22,** 1231–1240 (2010).

23. Meng, X., D'Arcy, C., Morgan, D. & Mousseau, D. D. Predicting the risk of dementia among Canadian seniors: a useable practice-friendly diagnostic algorithm. *Alzheimer Dis. Assoc. Disord.* **27,** 23–9 (2013).

24. Kivipelto, M. *et al.* Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study. *Lancet Neurol.* **5,** 735–741 (2006).

25. Reitz, C. *et al.* A Summary Risk Score for the Prediction of Alzheimer Disease in Elderly Persons. *Arch. Neurol.* **67,** 835–41 (2010).

26. Walters, K. *et al.* Predicting dementia risk in primary care: development and validation of the Dementia Risk Score using routinely collected data. *BMC Med.* **14,** 1–12 (2016).

27. Béland, Y. Canadian community health survey--methodological overview. *Heal. reports / Stat. Canada, Can. Cent. Heal. Inf. = Rapp. sur la sant?? / Stat. Canada, Cent. Can. d'information sur la sant??* **13,** 9–14 (2002).

28. Jaakkimainen, R. L. *et al.* Identification of Physician-Diagnosed Alzheimer's Disease and Related Dementias in Population-Based Administrative Data: A Validation Study Using Family Physicians' Electronic Medical Records. *J. Alzheimer's Dis.* **54,** 337–349 (2016).

29. Connolly, A., Gaehl, E., Martin, H., Morris, J. & Purandare, N. Underdiagnosis of

dementia in primary care: Variations in the observed prevalence and comparisons to the expected prevalence. *Aging Ment. Health* **15,** 978–984 (2011).

30. Kosteniuk, J. G. *et al.* Incidence and prevalence of dementia in linked administrative health data in Saskatchewan, Canada: a retrospective cohort study. *BMC Geriatr.* **15,** 73 (2015).

31. Harrell, F. E. *Regression Modeling Strategies with applicaitons to linear models, logistic regression and survival analysis*. (Springer, 2001).

32. Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. C. & Habbema, J. D. F. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J. Clin. Epidemiol.* **58,** 475–483 (2005).

33. Steyerberg, E. W. *Clinical Prediction Models*. (Springer, 2009).

34. Harrell, F. E. *Hmisc: Harrell Miscellaneous. R package version 4.0-2*. (2016).

35. R Core Team. R: A language and environment for statistical computing. (2016).

36. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur. Urol.* **67,** 1142–1151 (2015).

37. Exalto, L. G. *et al.* Midlife risk score for the prediction of dementia four decades later. *Alzheimer's Dement.* **10,** 562–570 (2014).

38. Exalto, L. G. *et al.* Risk score for prediction of 10 year dementia risk in individuals with type 2 diabetes: A cohort study. *Lancet Diabetes Endocrinol.* **1,** 183–190 (2013).

39. Taljaard, M. *et al.* Cardiovascular Disease Population Risk Tool (CVDPoRT): predictive algorithm for assessing CVD risk in the community setting. A study protocol. *BMJ Open* **4,** e006701 (2014).

40. Manuel, D. G. *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.* **13,** 1–27 (2016).

41. Pampalon, R., Hamel, D., Gamache, P. & Raymond, G. A deprivation index for health planning in Canada. *Chronic Dis. Can.* **29,** 178–91 (2009).

42. Rubin, D. B. & Schenker, N. Multiple imputation in health-care databases: an overview and some applications. *Stat. Med.* **10,** 585–98 (1991).

43. Therneau, T. M., Grambsch, P. M. & Fleming, T. R. Martingale-based residuals for survival models. *Biometrika* **77,** 147–160 (1990).

44. Allison, P. D. *Survival analysis using SAS : A practical guide*. (SAS Institute, 2010).

45. Van Houwelingen, J. C. & Le, C. S. Predictive value of statistical models. *Stat Med* **9,** 1303–1325 (1990).

46. Ambler, G., Brady, A. R. & Royston, P. Simplifying a prognostic model: A simulation study based on clinical data. *Stat. Med.* **21,** 3803–3822 (2002).

47. Nagelkerke, N. J. D. A Note on a General Definition of the Coefficient of Determination. *Biometrika* **78,** 691–692 (1991).

48. Arkes, H. R. *et al.* The covariance decomposition of the probability score and its use in evaluating prognostic estimates. SUPPORT Investigators. *Med. Decis. Making* **15,** 120–31 (1995).

49. Cook, N. R. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin. Chem.* **54,** 17–23 (2008).

50. Cook, N. R. Comment: Measures to summarize and compare the predictive capacity of markers. *Int. J. Biostat.* **6,** Article 22; discussion Article 25 (2010).

51. Krause, J., Perer, A. & Bertini, E. Using Visual Analytics to Interpret Predictive Machine Learning Models. *arXiv Prepr. arXiv1606.05685.* (2016).

Table 1. Pre-specification of predictor variables for DemPoRT with initial degrees of freedom (df) allocation

Variable	Scale	Initial Variable Specification	df
Socio-demographic Factors			
Age	Continuous	**5 knot spline:** Valid range: 55-102 (male), 55-101 (female)	4
Sex	Categorical	**Stratified:** Male; Female	NA
Ethnicity	Categorical	**7 categories:** 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
Education	Categorical	**4 categories:** Less than secondary school; Secondary school graduation; Some postsecondary; Postsecondary graduation	3
Marital Status	Categorical	**4 categories:** Now married/Common-law; Separated/Divorced; Widowed; Single	3
Neighborhood Social and Material Deprivation[41]	Ordinal	**3 categories:** Low (1st or 2nd quintile); High 4th or 5th quintile; Moderate (3rd quintile)	2
General Health			
Sense of belonging to local community	Ordinal	**4 categories:** Very strong; Somewhat strong; Somewhat weak; Very weak	3
Self-perceived stress	Ordinal	**5 categories:** Not at all stressful; Not very stressful; A bit stressful; Quite a bit stressful; Extremely stressful	4
Self-rated health	Ordinal	**5 categories:** Poor; Fair; Good; Very Good; Excellent	4
Health Behaviors			
Pack years of smoking	Continuous	**3 knot spline:** Valid range: 0-112 (male), 0-78 (female)	2
Smoking status	Categorical	**4 categories:** 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	**3 knot spline:** Valid range: 0-50 (male), 0-24 (female)	2
Former drinker	Dichotomous	Yes; No	1
Consumption of fruit, salad, carrot and other vegetables (average daily frequency)	Continuous	**3 knot spline:** Valid range: 0-48 (male), 0-31 (female)	2
Potato consumption (average daily frequency)	Continuous	**3 knot spline:** Valid range: 0-2	2
Juice consumption (average daily consumption	Continuous	**3 knot spline:** Valid range: 0-6 (male), 0-5 (female)	2
Leisure physical activity (average daily METs (kcal/kg/day))	Continuous	**3 knot spline:** Valid range: 0-16 (male), 0-12 (female)	2
Functional Measures			
Personal hygiene and care	Dichotomous	Does not need help; Needs help	1
Locomotion in the home	Dichotomous	Does not need help; Needs help	1
Meal preparation	Dichotomous	Does not need help; Needs help	1
Running errands	Dichotomous	Does not need help; Needs help	1
Ordinary housework	Dichotomous	Does not need help; Needs help	1
Heavy housework	Dichotomous	Does not need help; Needs help	1
Finances	Dichotomous	Does not need help; Needs help	1
Health Conditions			
Heart disease	Dichotomous	Yes; No	1
Stroke	Dichotomous	Yes; No	1
Diabetes	Dichotomous	Yes; No	1
Mood disorder	Dichotomous	Yes; No	1
High blood pressure	Dichotomous	Yes; No	1
Body mass index	Continuous	**3 knot spline:** Valid range: 10-44 (male), 10-47 (female)	2
Design			
Survey year	Ordinal	**4 categories:** 2000/01, 2002/03, 2004/05, 2006/07	3

DemPoRT, Dementia Population Risk Tool; df, degrees of freedom; MET, metabolic equivalent task

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	3
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	3
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	5-7
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	7,8
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	8,9
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	8-10
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	8,9
	5b	D;V	Describe eligibility criteria for participants.	8-10
	5c	D;V	Give details of treatments received, if relevant.	NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	9,10
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	12
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	11,12, 25
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	12
Sample size	8	D;V	Explain how the study size was arrived at.	8-12
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	12,13
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	13,14, 25
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	14,15
	10c	V	For validation, describe how the predictions were calculated.	16
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	16
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	16
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	16,17
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	10
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	NA
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	NA
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	NA
	15b	D	Explain how to the use the prediction model.	NA
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	NA
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	NA
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	4,17
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	NA
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	NA
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	5-7
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	4,18
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	19

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.

BMJ Open

Dementia Population Risk Tool (DemPoRT): Study Protocol for a Predictive Algorithm Assessing Dementia Risk in the Community

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018018.R2
Article Type:	Protocol
Date Submitted by the Author:	25-Aug-2017
Complete List of Authors:	Fisher, Stacey; Ottawa Hospital Research Institute; University of Ottawa, Department of Epidemiology and Community Medicine Hsu, Amy; Ottawa Hospital Research Institute; Institute of Clinical Evaluative Sciences Mojaverian, Nassim; Institute of Clinical Evaluative Sciences Taljaard, Monica; Ottawa Hospital Research Institute, Clinical Epidemiology Program; University of Ottawa, Department of Epidemiology and Community Medicine Huyer, Greg; University of Ottawa Telfer School of Management Manuel, Douglas; Ottawa Hospital Research Institute, Clinical Epidemiology Program; Institute of Clinical Evaluative Sciences Tanuseputro, Peter; Ottawa Hospital Research Institute, Clinical Epidemiology Program; Institute of Clinical Evaluative Sciences
<b>Primary Subject Heading</b>:	Neurology
Secondary Subject Heading:	Epidemiology, Geriatric medicine, Public health
Keywords:	Dementia < NEUROLOGY, Population projection, Risk stratification, Health behavior

Dementia Population Risk Tool (DemPoRT): Study Protocol for a Predictive Algorithm Assessing Dementia Risk in the Community

Stacey Fisher, MSc[1,2,3] stacey.fisher@uottawa.ca

Amy Hsu, PhD[1,2] ahsu@toh.ca

Nassim Mojaverian, MSc[2] namojaverian@ohri.ca

Monica Taljaard, PhD[1,3]mtaljaard@ohri.ca

Gregory Huyer, PhD[4] ghuye047@uottawa.ca

Doug Manuel, MD[1,2,3,5,6] dmanuel@ohri.ca

Peter Tanuseputro, MD[1,2,6,7,8] ptanuseputro@ohri.ca

[1] Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

[2] Institute for Clinical Evaluative Sciences, Ottawa, Ontario, Canada

[3] Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada

[4] Telfer School of Management, University of Ottawa, Ottawa, Ontario, Canada

[5] Statistics Canada, Ottawa, Ontario, Canada

[6] Department of Family Medicine, University of Ottawa, Ottawa, Ontario, Canada

[7] Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada

[8] Bruyère Research Institute, Ottawa, Ontario, Canada

Corresponding author:

Dr. Peter Tanuseputro

1053 Carling Ave.

Box 693

Ottawa, ON K1Y 4E9, Canada

ptanuseputro@ohri.ca

Phone: 613-798-5555

Word Count: 4 229

Keywords: Dementia; population projection; risk stratification; health behavior

ABSTRACT

Introduction: The burden of disease from dementia is a growing global concern as incidence increases dramatically with age and average life expectancy has been increasing around the world. Planning for an aging population requires reliable projections of dementia prevalence; however, existing population projections are simple and have poor predictive accuracy. The Dementia Population Risk Tool (DemPoRT) will predict incidence of dementia in the population setting using multivariable modeling techniques, and will be used to project dementia prevalence.

Methods and Analysis: The derivation cohort will consist of elderly Ontario respondents of the Canadian Community Health Survey (CCHS) (2001, 2003, 2005, 2007; 18 764 males and 25 288 females). Pre-specified predictors include sociodemographic, general health, behavioral, functional and health condition variables. Incident dementia will be identified through individual linkage of survey respondents to population-level administrative health care databases (1 797 and 3 281 events, and 117 795 and 166 573 person-years of follow-up, for males and females, respectively until March 31, 2014). Using time of first dementia capture as the primary outcome and death as a competing risk, sex-specific proportional hazards regression models will be estimated. The 2008/2009 CCHS survey will be used for validation (approximately 4 600 males and 6 300 females). Overall calibration and discrimination will be assessed as well as calibration within predefined subgroups of importance to clinicians and policy makers.

Ethics and Dissemination: Research ethics approval has been granted by the Ottawa Health Science Network Research Ethics Board. DemPoRT results will be submitted for publication in peer-review journals and presented at scientific meetings. The algorithm will be assessable online for both population and individual uses.

Trial Registration Number: ClinicalTrials.gov NCT03155815.

STRENGTHS AND LIMITATIONS

- The Dementia Population Risk Tool (DemPoRT) will be developed and validated using predictors from large population-based community health surveys that are individually linked to routinely-collected health administration data in Ontario. To our knowledge, DemPoRT will be the first algorithm designed to predict and project dementia incidence at the population-level.
- Although repeat predictor assessment and detailed cognitive testing to ascertain dementia diagnoses is preferable, it is not available or feasible at the population level.
- Statistical overfitting is a concern, however full pre-specification of the analysis plan and predictors will limit this risk.
- Although a rigorous approach to model development will be used, further validation will be needed to assess generalizability, and calibration will be required for application in other jurisdictions.
- DemPoRT will be used to produce improved estimates of future dementia burden, will assess the contribution of specific risk factors to the population risk, and will identify population subgroups at high risk of developing dementia. This information will be used by policymakers to prepare for and reduce dementia impact.

INTRODUCTION

The burden of disease from dementia is a growing global concern as incidence increases dramatically with age and average life expectancy has been increasing around the world[1,2]. Planning for an aging population requires reliable projections of dementia burden and the implications for resource requirements. Existing population-level projections for dementia, however, are overly simplistic and likely inaccurate[3].

Limitations of Current Dementia Projection Methodology

Almost all existing dementia projections have used extrapolation and macrosimulation methods, which are simplistic and make assumptions that may not hold true into the future[3]. Most extrapolations simply apply current age- and sex-specific prevalence estimates of dementia to future population projections. Macrosimulations typically use estimates of dementia incidence and mortality, stratified by age and sex, to simulate disease prevalence as the population ages[1,4–6]. Projections from extrapolations incorrectly assume that the risk of mortality among those with and without dementia are equivalent[7,8], and both methods assume that the age and sex-specific prevalence of dementia risk factors will not change with time. The assumption of stable risk factor prevalence is widely thought to be the major source of error in existing dementia projections[3,9–11].

Changing trends of dementia risk factors has the potential to have a dramatic impact on dementia prevalence estimates, as up to 50% of dementia cases have been attributed to modifiable factors[9,12], and the prevalence of several factors has been projected to change significantly in the near future. For example, the population prevalence of diabetes and obesity in Canada has been projected to increase, while smoking, hypertension and dyslipidemia have been projected to decline[13]. Consideration of risk factor prevalence is therefore important to improve the accuracy of dementia projections, and simple extrapolations and macrosimulations are often inadequate.

Predictive Multivariable Modeling of Dementia Incidence

Population-based **predictive risk algorithms** examine the effect of risk factors on dementia incidence, and can be used for dementia burden projection. Population-based data that contain detailed risk factor information, such as health surveys, are linked at the individual-level to administrative data that capture dementia development. A multivariable model of dementia incidence is derived, validated against external data, and predictive performance is assessed. Once developed, the algorithm can be used to project disease incidence and prevalence. To obtain prevalence projections, the algorithm can be integrated in to a microsimulation model such as Statistics Canada's Population Health Models (POHEM). POHEM dynamically models individual life trajectories of a population representative of Canada including births, deaths and migration, disease incidence and progression, and exposure to risk factors, facilitating detailed examination of the influence of changing risk factor prevalence on future dementia prevalence.

Predictive risk algorithms can also be used to describe the risk of dementia in the population, assess the contribution of specific risk factors to the population risk, identify high-risk groups, and evaluate risk reduction strategies.

Existing Dementia Prediction Models

Many models have been developed to predict risk of dementia[14–26], most with the primary goal of identifying individuals in the clinical setting at high risk. They have varying discriminative ability (c-statistics ranging from 0.49[16] to 0.89[17]) and have generally been derived from small samples, rarely including more than a few thousand individuals. Existing models are therefore simplistic, including few predictors and rarely including interaction or non-linear terms Existing models thus facilitates understanding and use by physicians in clinical practice, but limits discriminatory ability and predictive accuracy. Walters et al[26] developed an algorithm for predicting 5-year dementia risk among individuals 60-79 years of age in the United Kingdom using an enormous derivation dataset of 800 000 individuals, and a simple model. The derivation model had a c-statistic of 0.84 (95% CI: 0.81, 0.87), but a low positive predictive value at most risk thresholds, and therefore is poor at identifying those at high risk of dementia. Additionally, as most dementia risk models are intended for use in the clinical setting, many include results from neuropsycological tests[17–23], MRI findings[18] and APOE genotype[18,24,25]. The inclusion of these variables, however, limits the application of these models as these variables are not available at the population-level.

The objective of this study is to develop and validate the Dementia Population Risk Tool (DemPoRT) algorithm to predict dementia incidence in the population setting. This will be done using multivariable modelling techniques, linking self-reported risk factors captured by a population-based health survey in Canada with administrative databases across healthcare sectors that capture healthcare diagnosed dementia. DemPoRT will be developed with a using a large population-based dataset using only variables that are available at the population-level, allowing for population-level application. DemPoRT will also utilize many methodological improvements over existing models. This protocol pre-specifies the predictor variables and analytic plan for model development, reducing the potential for overfitting and bias, and improving transparency. Interaction terms and flexible functions for continuous predictors will be investigated, increasing potential discriminative ability. The pre-specified analytic plan avoids data-driven variable selection procedures, further reducing the potential for bias.

To our knowledge, the DemPoRT predictive model will be the first algorithm designed to predict and project dementia incidence at the population-level. It will be used to estimate the future burden of dementia using techniques that consider changes in risk factor prevalence and will identify modifiable risk factors that can be targeted by individuals, clinicians and policy makers to reduce the burden of dementia.

METHODS AND ANALYSIS

Study Design

Two DemPoRT models, one for males and females, will be derived and validated using population-based data in Ontario, Canada, a multicultural province with 13.6 million residents. Predictors will be obtained from the Canadian Community Health Surveys (CCHS), and outcomes (i.e., diagnosis of dementia) will be obtained from routinely-collected health care data.

The derivation cohorts will consist of eligible respondents of the 2001, 2003, 2005 and 2007 CCHS (Cycles 1.1, 2.1, 3.1 and 4.1), while validation cohorts will consist of respondents to the 2008/2009 cycle. The CCHS is a national, cross-sectional survey developed by Statistics Canada to collect information related to health and health care utilization of the Canadian population.

The survey has a multistage stratified cluster design that represents approximately 98% of the Canadian population aged 12 years and over and attained an average response rate of 79% over the study period. The CCHS is conducted through telephone and in-person interviews, and all responses are self-reported. The details of survey methodology have been published elsewhere[27]. Survey respondents will be excluded if they are less than 55 years of age at survey administration, self-reported a history of dementia, or are not eligible for Ontario's universal health insurance. If a respondent was included in more than one CCHS cycle, only their earliest survey response will be used.

Outcome

Survey respondents diagnosed with dementia will be identified through individual linkage to several population-based administrative databases at the Institute for Clinical Evaluative Sciences (ICES). Dementia case ascertainment is based on a validated definition: 1 hospital record OR 3 physician claim records at least 30 days apart within a 2-year period OR a dispensing record for a cholinesterase inhibitor from Ontario Drug Benefit (ODB). This definition has a sensitivity of 79.3% and a specificity of 99.1% when validated against emergency medical record (EMR) data[28]. Due to known underdiagnosis of dementia[29,30], we will supplement this definition by adding survey respondents with dementia codes captured on home care and long-term care assessments (dementia flag AND Cognitive Performance Scale [CPS] score ≥2) using the Resident Assessment Instrument-Home Care (RAI-HC) database and the Continuing Care Reporting System (CCRS), respectively. We have found this addition adds substantially (approximately 18%) to the number of dementia cases captured.

Survey respondents with dementia will be excluded if they meet the criteria for dementia within two years of survey administration (to remove potentially prevalent cases) or are younger than 65 years of age at the time of dementia diagnosis (to exclude early onset dementia which likely has a different set of risk factors). Eligible survey respondents will be followed from the date of survey administration or age 65, whichever came later, until the earliest date of: dementia ascertainment, death (defined as competing risk), loss to follow-up (defined as loss of healthcare eligibility) or end of study (March 31, 2014).

Sample Size

The male and female derivation cohorts consist of 18 764 and 25 288 respondents, and 117 795 and 166 573 person-years of follow-up, respectively. For predictive models with time to event outcomes the number of participants experiencing the event should exceed 10 times the number of degrees of freedom to ensure adequate sample size[31]. The number of dementia events in the derivation cohort is 1 797 for men and 3 281 for women; therefore, the maximum number of total degrees of freedom for each of the DemPoRT models is 179 and 328, respectively, which we do not anticipate surpassing.

The validation cohorts will consist of approximately 4 600 males and 6 300 females, and 15 000 and 21 000 person-years of follow-up, respectively. Vergouwe *et al*[32] recommend a minimum of 100 events and 100 non-events for external validation studies. We expect approximately 225 events for men and 400 for women in our validation cohort.

Analysis Plan

The analysis plan was developed following guidelines by Harrell[31] and Steyerberg[33] after accessing the derivation data set, but prior to model fitting or descriptive analyses involving exposure-outcome associations. This was done to avoid Type 1 error introduced by data-driven variable selection or model specification. Key considerations of our analysis approach include full pre-specification of the predictor variables, use of flexible functions for continuous predictors, and preserving statistical properties by avoiding data-driven variable selection procedures. Analysis will be conducted using Harrell's Hmisc[34] package of functions in R[35] as well as SAS v9.4.

This study protocol and the reporting of our model estimation results will be guided by the TRIPOD statement for multivariable predictive models[36].

Identification of Predictors

Predictor variables were identified through review of existing predictive algorithms for dementia[9,16,18–22,24–26,37,38] and comparison to available data collected in the CCHS. Variable inclusion was informed by consultation with subject-matter experts and the project's advisory team, and informed by our previous work developing predictive models for cardiovascular disease and life expectancy[39,40].

Variables with narrow distributions or insufficient variation were excluded. Obvious cases of redundancy (e.g. alternate definitions of the same underlying behavior) were not included. A total of 32 predictor variables were identified: 7 sociodemographic, 3 general health, 9 behavioral, 7 functional, 5 health conditions and 1 design variable (CCHS survey cycle). As the effect of dementia risk factors varies by sex, separate models will be derived for men and women. Education, rather than individual income, was selected as a predictor due to several concerns with income including lack of generalizability, measurement error, stability over time and substantial missing values. Neighborhood social and material deprivation is captured using Pampalon's deprivation index[41]. Indicator variables for smoking status were created to allow the inclusion of smoking pack-years as a continuous predictor. The models will additionally include age interactions with the behavioral, functional and health condition variables as the effect of these risk factors on dementia are expected to vary with age. Detailed definitions and measurement of the predictor variables are presented in Table 1.

Data Cleaning and Coding of Predictors

Continuous variables will be inspected using boxplots and descriptive statistics to determine values outside a plausible range. Values that are clearly erroneous will be corrected, where possible, or set to missing. Continuous predictors with highly skewed distributions will be truncated to the 99.5th percentile. Categorization of continuous variables will be avoided to minimize the loss of predictive information. All data cleaning and coding will occur prior to examining exposure-outcome associations.

Missing Data

As traditional complete cases analyses suffer from inefficiency, selection bias, and other limitations[33], multiple imputation methods will be used to impute missing values using the 'aregImpute' function in the HMisc library[34]. This function simultaneously imputes missing values using predictive mean matching and uses bootstrapping to take all aspects of uncertainty in to account. The imputation model will consist of the full list of predictor variables, time to event and censoring variables, as well as auxiliary variables that are not predictors, but may nevertheless be useful in generating imputed values (e.g., income). The final model will be estimated in each of five multiple imputation data sets and the results combined using the rules developed by Rubin and Schenker[42] to account for imputation uncertainty.

Model Specification

Initial sex-specific main effects models will be fit using the pre-specified predictors and an initial degree of freedom allocation for each predictor (Table 1). Decisions on initial degree of freedom allocations were informed by the anticipated importance of each predictor and known dose-response relationships with dementia. Continuous predictors will be flexibly modelled using restricted cubic splines, with the knots placed at fixed quantiles of the distribution (e.g., 5th, 27.5th, 50th, 72.5th, and 95th centiles). Frequency distributions for categorical predictors will be examined and categories with small numbers of respondents will be combined, with analysts blinded to the number of events per category, to avoid instability in the regression analyses. Ordinal variables will be specified as either linear terms or as categorical if the expected association is more complex. Interactions will be restricted to linear terms. The initial model specification, presented in Table 1, includes a total of 86 degrees of freedom (63 main, 23 interaction).

Partial association chi-square statistics for each predictor 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 retain their initial degrees of freedom, while predictors with lower predictive potential will be modeled as simple linear terms or recoded by combining infrequent categories. This process of model specification does not increase the Type I error rate because all predictors will be retained in the full model regardless of their strength of association[31].

Model Estimation

The initial models will be estimated using competing risk Cox proportional hazards regression with time to dementia ascertainment as the outcome and death as a competing risk. Alternative model specifications, including subdistribution hazard and flexible parametric models, will be considered. All predictors will be centered about their means. A formal check of multicollinearity will be carried out using a variable clustering algorithm[31].

Proportional hazards models assume that the relative risk of the outcome between strata of predictors and the baseline risk must be constant over time. Violation of this assumption has been shown to produce biased results[43] although it has also been argued that the estimated coefficients of time-varying variables can simply be interpreted as an average rather than instantaneous hazard[44]. Plots of raw and smoothed scaled Schoenfeld residuals versus time for each predictor will be assessed to test this assumption and identify non-proportionality. If a violation of this assumption is identified we will consider addition of interaction terms between the predictor and log-transformed time.

Although the risk of overfitting will be minimal due to pre-specification of the models and a large sample size, the need for overfitting adjustment will be assessed. The degree of overfitting will be estimated using the heuristic shrinkage estimator, based on the log likelihood ratio chi-square statistic for the full model[45]. If shrinkage is <0.90, models will be adjusted for overfitting.

Estimation of the Reduced Models

Model pre-specification has advantages in limiting overfitting and spurious statistical significance but can result in a final model that is overly complex, difficult to interpret, and difficult to apply. Unnecessary predictor variables also distort the estimated effects of other predictors making the model more computationally intensive. It is suggested that a more parsimonious model that retains most of the prognostic information and performs as well as or better than the full model can be derived without increasing the Type 1 error rate[31,46]. We will identify a more parsimonious model using a stepdown procedure described by Ambler[46], which involves deleting the variable that results in the smallest decrease in model R^2 until removal leads to an R^2 that is less than a desired level. The reduced model will be evaluated against the full model using Akaike's Information Criterion, and by examining the effect on discrimination and calibration.

DemPoRT will be developed and validated using temporal split samples, however the final regression coefficients will use the full data set to maximize follow-up duration. A cohort-specific intercept and/or interaction term may be included in the final model if the derivation and validation cohorts differ; otherwise, the final combined model will maintain the same predictors and form as the derivation model.

Assessment of Predictive Performance

Predictive performance in the derivation and validation cohorts will be assessed and reported using overall measures of predictive accuracy, discrimination and calibration. Accuracy will be assessed with Nagelkerke's R^2 [47] and the Brier score[48]. Discrimination will be assessed using the concordance statistic. Model calibration is especially important in the development of prognostic models, as probabilities of future risk are of primary interest[33,49,50]. Calibration will be assessed by comparing the observed and predicted risk of dementia within vigintiles (20 groups of equal frequency) of predicted risk with emphasis on visual inspection of plots rather than formal statistical significance testing, which can be influenced by large sample sizes[32]. Calibration slopes will be generated by regressing the outcome in the validation cohort on the predicted dementia risk, reflecting the combined effect of overfitting to the derivation data as well as true differences in effects of predictors. Deviation of the slope from 1 (perfect calibration) will be tested using a Wald or likelihood ratio test. Calibration within predefined subgroups of importance to clinicians and policy makers (e.g., age group, health behavior, sociodemographic groups and health conditions) will additionally be evaluated. The clinically relevant standard of calibration was defined as less than 20% difference between observed and predicted estimates within subgroups with a dementia prevalence of at least 5%. All model performance measures will be calculated using the first of the multiply imputed data sets.

Model Presentation

The final regression model, derived from the combined sample of the derivation and validation cohorts, will be presented using estimated hazard ratios and 95% confidence intervals, along with results for the derivation and validation cohorts separately. We have found, however, this usual presentation less meaningful when presenting complex models[39]. To allow interpretation of the estimated effect of each predictor, model behavior will additionally be described using interactive visual tools to display the shape of the effect of each predictor[51]. The regression formula will also be published and used as the basis for web-based implementation.

Analyses Beyond Initial Model Development

We will conduct further analyses exploring the added predictive ability of novel risk factors that were ascertained in single CCHS cycles (e.g. sedentary activity, cognitive stimulation, sleep quality and duration, deafness), as well as risk factors that can be ascertained through linkage of additional data sources and similar cohorts (e.g. air pollution, detailed dietary consumption, lipid levels, blood pressure). In addition, sensitivity analysis of the age at survey administration cutoff used for cohort creation will be performed.

Once developed, DemPoRT will be used to project dementia incidence under different assumptions by entering counterfactual risk factor levels in to the algorithm at the population level, or at individual level and summed, and will be integrated in to POHEM for microsimulation modelling of prevalence projections.

A second, causal model (DemPoRT-C) will also be created to assess the relative contribution of lifestyle, socio-demographic and health factors to dementia incidence. Development will exclude variables believed to be in the causal pathway of dementia occurrence (e.g., self-rated health and functional measures) to reduce the attenuation of hazards from upstream risk factors, but will otherwise be the same as in the predictive model. DemPoRT-C will be applied to the most recent unlinked national CCHS survey.

LIMITATIONS

One of the limitations of this study will be the potential for misclassification error resulting from the use of self-reported predictors captured at one point in time and administrative data for outcome ascertainment. However, discriminating and well-calibrated algorithms have been developed using self-report information and although detailed cognitive testing to ascertain dementia diagnoses is preferable over the use of administrative data, it is not available or feasible at the population level. Another concern common to the development of highly complex risk algorithms such as DemPoRT, is the potential for statistical overfitting and increased Type 1 error, which can occur when the relationship between a predictor and the outcome influences whether it is used, and how it is fit. This risk is reduced by pre-specification of the predictors and analytic plan, as we have done in this protocol. The model will also be adjusted for overfitting if necessary, as specified previously. Lastly, although a rigorous approach to model development will be used, further validation will be needed to assess generalizability, and calibration will be required for application in other jurisdictions.

ETHICS AND MODEL DISSEMINATION

The DemPoRT project advisory committee has been created to ensure that the models meet the needs of knowledge users. This committee has worked with the study team to identify predictors of dementia based on scientific and policy importance and will aid in the identification of important target populations and the establishment of policy-relevant differences for calibration studies.

DemPoRT results will be submitted for publication in peer-review journals and presented at scientific meetings. A web-based individual-level calculator will be created if the models are appropriate for individual use. Although DemPoRT emphasizes risk prediction at the population-level, we have found that individual-level calculators are an effective engagement and translation tool for both the general public and knowledge users.

CONCLUSIONS

To the best of our knowledge, DemPoRT will be the first population-based algorithm designed to predicting and projecting dementia incidence at the population level. The DemPoRT models will produce estimates of future dementia burden that we believe will be more accurate than existing estimates, will assess the contribution of specific risk factors to the population risk, and identify groups at high risk of developing dementia. Although a rigorous approach to model development will be used, further validation will be needed to assess generalizability, and calibration will be required for application in other jurisdictions.

CONTRIBUTIONS

SF drafted and revised the manuscript, and contributed to study design and protocol development. NM contributed to study design, protocol development and provided data/statistical support. AH, MT, DM and GH contributed to the design of the study and protocol development. PT is the lead investigator of the study and was responsible for the conception of the project, the grant application, study design and protocol development. All authors provided critical reviews of the manuscript and approved the final version.

FUNDING

This work is supported by the Canadian Institutes of Health Research operating grant MOP 142237. SF is supported by a doctoral award from the Canadian Institutes of Health Research.

ACKNOWLEDGEMENTS

This project is supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The sponsors have no role in the design or conduct of the study; in the collection, analysis or interpretation of the data; or in the preparation, review or approval or the manuscript. The options, results and conclusions reported are those of the authors and are independent from the funding source. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred.

COMPETING INTERESTS

None declared.

ETHICS APPROVAL

Research ethics approval has been granted by the Ottawa Health Science Network Research Ethics Board.

REFERNCES

1. Brookmeyer, R., Gray, S. & Kawas, C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. *Am. J. Public Health* **88,** 1337–1342 (1998).
2. Brayne, C. The elephant in the room - healthy brains in later life, epidemiology and public health. *Nat. Rev. Neurosci.* **8,** 233–9 (2007).
3. Norton, S., Matthews, F. E. & Brayne, C. A commentary on studies presenting projections of the future prevalence of dementia. *BMC Public Health* **13,** 1 (2013).
4. Sloane, P. D. *et al.* The public health impact of Alzheimer's disease, 2000-2050: potential implication of treatment advances. *Annu. Rev. Public Health* **23,** 213–31 (2002).
5. Mura, T., Dartigues, J. F. & Berr, C. How many dementia cases in France and Europe? Alternative projections and scenarios 2010-2050. *Eur. J. Neurol.* **17,** 252–259 (2010).
6. Hebert, L. E., Scherr, P. A., Bienias, J. L., Bennett, D. A. & Evans, D. A. Alzheimer disease in the US population: prevalence estimates using the 2000 census. *Arch. Neurol.* **60,** 1119–22 (2003).
7. Brookmeyer, R., Johnson, E., Ziegler-Graham, K. & Arrighi, H. M. Forecasting the global burden of Alzheimer's disease. *Alzheimer's Dement.* **3,** 186–191 (2007).
8. Dewey, M. E. & Chen, C.-M. Neurosis and mortality in persons aged 65 and over living in the community: a systematic review of the literature. *Int. J. Geriatr. Psychiatry* **19,** 554–7 (2004).
9. Norton, S., Matthews, F. E., Barnes, D. E., Yaffe, K. & Brayne, C. Potential for primary prevention of Alzheimer's disease: An analysis of population-based data. *Lancet Neurol.* **13,** 788–794 (2014).

10. Joly, P. *et al.* Prevalence projections of chronic diseases and impact of public health intervention. *Biometrics* **69,** 109–17 (2013).

11. Lee, Y. The recent decline in prevalence of dementia in developed countries: implications for prevention in the Republic of Korea. *J. Korean Med. Sci.* **29,** 913–8 (2014).

12. Barnes, D. E. & Yaffe, K. The projected effect of risk factor reduction on Alzheimer's disease prevalence. *Lancet. Neurol.* **10,** 819–28 (2011).

13. Manuel, D. G. *et al.* Projections of preventable risks for cardiovascular disease in Canada to 2021: a microsimulation modelling approach. *C. Open* **2,** E94–E101 (2014).

14. Tang, E. Y. H. *et al.* Current developments in dementia risk prediction modelling: An updated systematic review. *PLoS One* **10,** 1–31 (2015).

15. Stephan, B. C. M., Kurth, T., Matthews, F. E., Brayne, C. & Dufouil, C. Dementia risk prediction in the population: are screening models accurate? *Nat. Rev. Neurol.* **6,** 318–326 (2010).

16. Anstey, K. J. *et al.* A self-report risk index to predict occurrence of dementia in three independent cohorts of older adults: The ANU-ADRI. *PLoS One* **9,** (2014).

17. Wolfsgruber, S. *et al.* The CERAD neuropsychological assessment battery total score detects and predicts alzheimer disease dementia with high diagnostic accuracy. *Am. J. Geriatr. Psychiatry* **22,** 1017–1028 (2014).

18. Barnes, D. E. *et al.* Predicting risk of dementia in older adults: The late-life dementia risk index. *Neurology* **73,** 173–179 (2009).

19. Chary, E. *et al.* Short- versus long-term prediction of dementia among subjects with low and high educational levels. *Alzheimers. Dement.* **9,** 562–71 (2013).

20. Jessen, F. *et al.* Prediction of Dementia in Primary Care Patients. *PLoS One* **6,** e16852

(2011).

21. Song, X., Mitnitski, A. & Rockwood, K. Nontraditional risk factors combine to predict Alzheimer disease and dementia. *Neurology* **77,** 227–234 (2011).

22. Tierney, M. C., Moineddin, R. & McDowell, I. Prediction of all-cause dementia using neuropsychological tests within 10 and 5 years of diagnosis in a community-based sample. *J. Alzheimer's Dis.* **22,** 1231–1240 (2010).

23. Meng, X., D'Arcy, C., Morgan, D. & Mousseau, D. D. Predicting the risk of dementia among Canadian seniors: a useable practice-friendly diagnostic algorithm. *Alzheimer Dis. Assoc. Disord.* **27,** 23–9 (2013).

24. Kivipelto, M. *et al.* Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study. *Lancet Neurol.* **5,** 735–741 (2006).

25. Reitz, C. *et al.* A Summary Risk Score for the Prediction of Alzheimer Disease in Elderly Persons. *Arch. Neurol.* **67,** 835–41 (2010).

26. Walters, K. *et al.* Predicting dementia risk in primary care: development and validation of the Dementia Risk Score using routinely collected data. *BMC Med.* **14,** 1–12 (2016).

27. Béland, Y. Canadian community health survey--methodological overview. *Heal. reports / Stat. Canada, Can. Cent. Heal. Inf. = Rapp. sur la sant?? / Stat. Canada, Cent. Can. d'information sur la sant??* **13,** 9–14 (2002).

28. Jaakkimainen, R. L. *et al.* Identification of Physician-Diagnosed Alzheimer's Disease and Related Dementias in Population-Based Administrative Data: A Validation Study Using Family Physicians' Electronic Medical Records. *J. Alzheimer's Dis.* **54,** 337–349 (2016).

29. Connolly, A., Gaehl, E., Martin, H., Morris, J. & Purandare, N. Underdiagnosis of

dementia in primary care: Variations in the observed prevalence and comparisons to the expected prevalence. *Aging Ment. Health* **15,** 978–984 (2011).

30. Kosteniuk, J. G. *et al.* Incidence and prevalence of dementia in linked administrative health data in Saskatchewan, Canada: a retrospective cohort study. *BMC Geriatr.* **15,** 73 (2015).

31. Harrell, F. E. *Regression Modeling Strategies with applicaitons to linear models, logistic regression and survival analysis*. (Springer, 2001).

32. Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. C. & Habbema, J. D. F. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J. Clin. Epidemiol.* **58,** 475–483 (2005).

33. Steyerberg, E. W. *Clinical Prediction Models*. (Springer, 2009).

34. Harrell, F. E. *Hmisc: Harrell Miscellaneous. R package version 4.0-2*. (2016).

35. R Core Team. R: A language and environment for statistical computing. (2016).

36. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur. Urol.* **67,** 1142–1151 (2015).

37. Exalto, L. G. *et al.* Midlife risk score for the prediction of dementia four decades later. *Alzheimer's Dement.* **10,** 562–570 (2014).

38. Exalto, L. G. *et al.* Risk score for prediction of 10 year dementia risk in individuals with type 2 diabetes: A cohort study. *Lancet Diabetes Endocrinol.* **1,** 183–190 (2013).

39. Taljaard, M. *et al.* Cardiovascular Disease Population Risk Tool (CVDPoRT): predictive algorithm for assessing CVD risk in the community setting. A study protocol. *BMJ Open* **4,** e006701 (2014).

40. Manuel, D. G. *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.* **13,** 1–27 (2016).

41. Pampalon, R., Hamel, D., Gamache, P. & Raymond, G. A deprivation index for health planning in Canada. *Chronic Dis. Can.* **29,** 178–91 (2009).

42. Rubin, D. B. & Schenker, N. Multiple imputation in health-care databases: an overview and some applications. *Stat. Med.* **10,** 585–98 (1991).

43. Therneau, T. M., Grambsch, P. M. & Fleming, T. R. Martingale-based residuals for survival models. *Biometrika* **77,** 147–160 (1990).

44. Allison, P. D. *Survival analysis using SAS : A practical guide*. (SAS Institute, 2010).

45. Van Houwelingen, J. C. & Le, C. S. Predictive value of statistical models. *Stat Med* **9,** 1303–1325 (1990).

46. Ambler, G., Brady, A. R. & Royston, P. Simplifying a prognostic model: A simulation study based on clinical data. *Stat. Med.* **21,** 3803–3822 (2002).

47. Nagelkerke, N. J. D. A Note on a General Definition of the Coefficient of Determination. *Biometrika* **78,** 691–692 (1991).

48. Arkes, H. R. *et al.* The covariance decomposition of the probability score and its use in evaluating prognostic estimates. SUPPORT Investigators. *Med. Decis. Making* **15,** 120–31 (1995).

49. Cook, N. R. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin. Chem.* **54,** 17–23 (2008).

50. Cook, N. R. Comment: Measures to summarize and compare the predictive capacity of markers. *Int. J. Biostat.* **6,** Article 22; discussion Article 25 (2010).

51. Krause, J., Perer, A. & Bertini, E. Using Visual Analytics to Interpret Predictive Machine Learning Models. *arXiv Prepr. arXiv1606.05685.* (2016).

Table 1. Pre-specification of predictor variables for DemPoRT with initial degrees of freedom (df) allocation

Variable	Scale	Initial Variable Specification	df
Socio-demographic Factors			
Age	Continuous	**5 knot spline:** Valid range: 55-102 (male), 55-101 (female)	4
Sex	Categorical	**Stratified:** Male; Female	NA
Ethnicity	Categorical	**7 categories:** 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
Education	Categorical	**4 categories:** Less than secondary school; Secondary school graduation; Some postsecondary; Postsecondary graduation	3
Marital Status	Categorical	**4 categories:** Now married/Common-law; Separated/Divorced; Widowed; Single	3
Neighborhood Social and Material Deprivation[41]	Ordinal	**3 categories:** Low (1st or 2nd quintile); High 4th or 5th quintile; Moderate (3rd quintile)	2
General Health			
Sense of belonging to local community	Ordinal	**4 categories:** Very strong; Somewhat strong; Somewhat weak; Very weak	3
Self-perceived stress	Ordinal	**5 categories:** Not at all stressful; Not very stressful; A bit stressful; Quite a bit stressful; Extremely stressful	4
Self-rated health	Ordinal	**5 categories:** Poor; Fair; Good; Very Good; Excellent	4
Health Behaviors			
Pack years of smoking	Continuous	**3 knot spline:** Valid range: 0-112 (male), 0-78 (female)	2
Smoking status	Categorical	**4 categories:** 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	**3 knot spline:** Valid range: 0-50 (male), 0-24 (female)	2
Former drinker	Dichotomous	Yes; No	1
Consumption of fruit, salad, carrot and other vegetables (average daily frequency)	Continuous	**3 knot spline:** Valid range: 0-48 (male), 0-31 (female)	2
Potato consumption (average daily frequency)	Continuous	**3 knot spline:** Valid range: 0-2	2
Juice consumption (average daily consumption	Continuous	**3 knot spline:** Valid range: 0-6 (male), 0-5 (female)	2
Leisure physical activity (average daily METs (kcal/kg/day))	Continuous	**3 knot spline:** Valid range: 0-16 (male), 0-12 (female)	2
Functional Measures			
Personal hygiene and care	Dichotomous	Does not need help; Needs help	1
Locomotion in the home	Dichotomous	Does not need help; Needs help	1
Meal preparation	Dichotomous	Does not need help; Needs help	1
Running errands	Dichotomous	Does not need help; Needs help	1
Ordinary housework	Dichotomous	Does not need help; Needs help	1
Heavy housework	Dichotomous	Does not need help; Needs help	1
Finances	Dichotomous	Does not need help; Needs help	1
Health Conditions			
Heart disease	Dichotomous	Yes; No	1
Stroke	Dichotomous	Yes; No	1
Diabetes	Dichotomous	Yes; No	1
Mood disorder	Dichotomous	Yes; No	1
High blood pressure	Dichotomous	Yes; No	1
Body mass index	Continuous	**3 knot spline:** Valid range: 10-44 (male), 10-47 (female)	2
Design			
Survey year	Ordinal	**4 categories:** 2000/01, 2002/03, 2004/05, 2006/07	3

DemPoRT, Dementia Population Risk Tool; df, degrees of freedom; MET, metabolic equivalent task

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	3
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	3
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	5-7
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	7,8
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	8,9
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	8-10
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	8,9
	5b	D;V	Describe eligibility criteria for participants.	8-10
	5c	D;V	Give details of treatments received, if relevant.	NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	9,10
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	12
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	11,12, 25
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	12
Sample size	8	D;V	Explain how the study size was arrived at.	8-12
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	12,13
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	13,14, 25
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	14,15
	10c	V	For validation, describe how the predictions were calculated.	16
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	16
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	16
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	16,17
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	10
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	NA
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	NA
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	NA
	15b	D	Explain how to the use the prediction model.	NA
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	NA
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	NA
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	4,17
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	NA
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	NA
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	5-7
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	4,18/ 19
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	20

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.