Population-Based Risk Prediction Models to Predict and Prevent Premature Mortality in Canadian Cities

Lief Pagalan[†,*]
[†] Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

Keywords: population health, premature mortality, precision public health, machine learning, social determinants of health

1. Introduction

Health system planners are increasingly interested in using population-level data to inform system planning and population-level health interventions.[1–4] Precision public health offers to improve the health system's ability to predict and prevent public health risks by developing policies and targeted public health interventions aimed at specific, high-risk subpopulations.[5, 6] Premature mortality, defined as deaths before the age of 75, is used as an indicator to assess how a health system is functioning and to compare health status between groups, regions, and health systems. Moreover, premature deaths are preventable through policy, effective public health interventions, and timely medical treatment.[7, 8] In Canada, premature mortality has remained stagnant, but gaps in premature mortality have been growing across sex, socioeconomic status, and geography.[9–12] Such variations in premature mortality indicate that health systems are not functioning equitably.

Past predictive modelling in health research has focused on clinical decision-making in individual patients or clinical subgroups.[13–18] To date, little work has been done to develop risk prediction models for population and public health, and few studies have focused explicitly on all-cause premature mortality. Our study will be among the first to incorporate both robust measures of social determinants of health and environmental risk factors (e.g., neighbourhood deprivation, air pollution concentrations, proximity to green space, and built environment characteristics like urban density, walkability, and land use) into prediction models. Including social and environmental data into population health prediction models recognizes upstream factors and can improve predictive performance since living conditions, neighbourhoods, and community characteristics have significant health impacts.

2. Objectives

We propose developing and testing a population-based risk prediction model to predict the five-year incidence of all-cause premature mortality in Canadian cities using machine learning methods. We will train and test the risk prediction model on a representative sample of the Canadian population, and we will incorporate individual and neighbourhood-level data on social and environmental determinants of health. Using representative data of the population and integrating social and environmental risk factors, we expect to enhance model predictive accuracy and fairness while equitably addressing the growing social gradients in premature mortality. By incorporating social and environmental determinants and area-level variables, we aim to capture residual unfairness from structural and social factors that are not captured at the individual level but still shape individual outcomes.[19]

[*] lief.pagalan@utoronto.ca

Working alongside our health system partners, we will test and deploy the finalized risk prediction model to create population risk segments and identify and describe high-risk sub-populations. Our validated risk prediction model will support population health management and help inform policies and social and environmental interventions that can minimize health care expenditures and ensure healthy living conditions for all.

3. Methods

We will link sociodemographic, health behaviour, and mortality data from the Canadian Community Health Survey (CCHS) and the Vital Statistics Death Database (CVSD) to urban, environmental, and land use data.[20–24] The data will contain upwards of 400 sociodemographic and health behaviour features and approximately 100 environmental ones. The CCHS is a cross-sectional survey representing approximately 98% of the Canadian population and collects information on health status, health care utilization, and health determinants among Canadians ages 12 and older.[25] Urban and environmental data will include health inequity, active living friendliness, green space, air pollution, weather, and land use features.[26–32] Survey respondents will be excluded if they were under 18 or older than 70 years of age as of the CCHS interview date. A cut-off of 70 years of age enables a consistent five-year follow-up period. Survey respondents will be further restricted to individuals who resided in Canadian census metropolitan areas (CMAs). After exclusions, we expect a sample size of approximately 300,000 individuals and 72,000 premature deaths.

Machine learning methods will enable us to model premature mortality as a health outcome with complex and multifactorial pathways, which is not possible with traditional epidemiological approaches. We will test penalized logistic regression with engineered features to capture non-additive and interaction effects and extreme gradient boosting models, both supervised learning, binary classification methods, and compare their performance. K-fold cross-validation will be used to train and validate the models with 70:20:10 training, validation, and test splits. We will report Brier scores for overall model performance; sensitivity, specificity, c-statistics, F1-scores, and ROC and precision-recall-gain curve plots for discrimination; and calibration plots for goodness-of-fit.

To assess the impact of area-level social determinants and environmental features, we will compare the predictive performance between the original model versus models developed using (a.) only the CCHS, (b.) the CCHS and area-level social determinant, and (c.) the CCHS and environmental exposures. We will also evaluate group fairness during algorithm development by assessing whether the model's predictive performance and calibration are equal across age groups, ethnicities, and income quintiles. Since the risk prediction model is being developed for a range of urban centres across Canada, we will also evaluate model performance across CMAs by Canadian regions (i.e., Atlantic, Quebec, Ontario, Prairies, British Columbia, Territories) and between small, mid-sized, and large CMAs.

4. Progress to Date

CCHS and CVSD data linkages were completed in December 2020. Linkages to area-level social determinants and environmental exposures are expected to be completed by Summer 2021. Preliminary model development is scheduled to begin in Fall 2021.

Acknowledgements

Lief Pagalan is supervised by Dr. Laura C. Rosella, Dr. Marzyeh Ghassemi, and Dr. Hong Chen, with research support from Meghan O'Neill, Mack Hurst, Lori Diemert, and Dr. Stacey Fisher. He is supported by a Frederick Banting and Charles Best Canada Graduate Scholarship Doctoral Award from the Canadian Institutes of Health Research.

References

[1] L. C. Rosella, C. Bornbaum, K. Kornas, M. Lebenbaum, L. Peirson, R. Fransoo, C. Loeppky, C. Gardner, and D. Mowat. "Evaluating the Process and Outcomes of a Knowledge Translation Approach to Supporting Use of the Diabetes Population Risk Tool (DPoRT) in Public Health Practice". In: *Canadian Journal of Program Evaluation* 33.1 (June 2018). ISSN: 1496-7308, 0834-1516. DOI: 10/ggkt3f. URL: https://journalhosting.ucalgary.ca/index.php/cjpe/article/view/31160 (visited on 02/14/2020).

[2] R. Ng, R. Sutradhar, W. P. Wodchis, and L. C. Rosella. "Chronic Disease Population Risk Tool (CDPoRT): a study protocol for a prediction model that assesses population-based chronic disease incidence". en. In: *Diagnostic and Prognostic Research* 2.1 (Dec. 2018), p. 19. ISSN: 2397-7523. DOI: 10/ggh8wd. URL: https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-018-0042-5 (visited on 01/21/2020).

[3] J. D. Morgenstern, L. C. Rosella, M. J. Daley, V. Goel, H. J. Schünemann, and T. Piggott. "'AI's gonna have an impact on everything in society, so it has to have an impact on public health': a fundamental qualitative descriptive study of the implications of artificial intelligence for public health". In: *BMC Public Health* 21.1 (Jan. 2021), p. 40. ISSN: 1471-2458. DOI: 10/ghw5jc. URL: https://doi.org/10.1186/s12889-020-10030-x (visited on 02/02/2021).

[4] J. D. Morgenstern, E. Buajitti, M. O'Neill, T. Piggott, V. Goel, D. Fridman, K. Kornas, and L. C. Rosella. "Predicting population health with machine learning: a scoping review". eng. In: *BMJ open* 10.10 (Oct. 2020), e037860. ISSN: 2044-6055. DOI: 10/ghhtwr.

[5] M. Dobbins and D. Buckeridge. "Precision public health: Dream or reality?" In: *Canada Communicable Disease Report* (June 2020), pp. 160–160. ISSN: 14818531. DOI: 10/gh69cs. URL: https://www.canada.ca/content/dam/phac-aspc/documents/services/reports-publications/canada-communicable-disease-report-ccdr/monthly-issue/2020-46/issue-6-june-4-2020/ccdrv46i06a01-eng.pdf (visited on 03/05/2021).

[6] M. Prosperi, J. S. Min, J. Bian, and F. Modave. "Big data hurdles in precision medicine and precision public health". In: *BMC Medical Informatics and Decision Making* 18.1 (Dec. 2018), p. 139. ISSN: 1472-6947. DOI: 10/gg6s3w. URL: https://doi.org/10.1186/s12911-018-0719-2 (visited on 07/20/2020).

[7] Canadian Institute for Health Information. *Health Indicators 2012*. Tech. rep. Ottawa, ON: Canadian Institute for Health Information, 2012. URL: https://secure.cihi.ca/free_products/health_indicators_2012_en.pdf.

[8] P. L. Remington, B. B. Catlin, and D. A. Kindig. "Monitoring progress in population health: trends in premature death rates". eng. In: *Preventing Chronic Disease* 10 (Dec. 2013), E214. ISSN: 1545-1151. DOI: 10/gbfgvk.

[9] Statistics Canada. *Premature and potentially avoidable mortality, three-year period, Canada, provinces, territories, health regions and peer groups*. 2020. URL: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310074301 (visited on 02/03/2020).

[10] E. Buajitti, T. Watson, K. Kornas, C. C. Bornbaum, D. Henry, and L. Rosella. "Ontario atlas of adult mortality, 1992-2015: Trends in Local Health Integration Networks. Toronto, ON: Population Health Analytics Lab". en. In: (2018). DOI: 10/gh7b7t. URL: http://rgdoi.net/10.13140/RG.2.2.21994.21440 (visited on 03/06/2021).

[11] E. Buajitti, T. Watson, T. Norwood, K. Kornas, C. Bornbaum, D. Henry, and L. C. Rosella. "Regional variation of premature mortality in Ontario, Canada: a spatial analysis". en. In: *Population Health Metrics* 17.1 (Dec. 2019), p. 9. ISSN: 1478-7954. DOI: 10/ggh8wc. URL: https://pophealthmetrics.biomedcentral.com/articles/10.1186/s12963-019-0193-9 (visited on 01/21/2020).

[12] F. V. Shahidi, A. Parnia, and A. Siddiqi. "Trends in socioeconomic inequalities in premature and avoidable mortality in Canada, 1991–2016". en. In: *CMAJ* 192.39 (Sept. 2020), E1114–E1128. ISSN: 0820-3946, 1488-2329. DOI: 10/ghfj54. URL: https://www.cmaj.ca/content/192/39/E1114 (visited on 10/13/2020).

[13] A. Abbasi, L. M. Peelen, E. Corpeleijn, Y. T. v. d. Schouw, R. P. Stolk, A. M. W. Spijkerman, D. L. v. d. A, K. G. M. Moons, G. Navis, S. J. L. Bakker, and J. W. J. Beulens. "Prediction models for risk of developing type 2 diabetes: systematic literature search and independent

external validation study". en. In: *BMJ* 345 (Sept. 2012), e5900. ISSN: 1756-1833. DOI: 10/gb3sdd. URL: https://www.bmj.com/content/345/bmj.e5900 (visited on 03/06/2021).

[14] P Brindle, A Beswick, T Fahey, and S Ebrahim. "Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review". en. In: *Heart* 92.12 (Dec. 2006), pp. 1752–1759. ISSN: 1355-6037. DOI: 10/fcdbfs. URL: https://heart.bmj.com/lookup/doi/10.1136/hrt.2006.087932 (visited on 03/06/2021).

[15] C. Meads, I. Ahmed, and R. D. Riley. "A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance". en. In: *Breast Cancer Research and Treatment* 132.2 (Apr. 2012), pp. 365–377. ISSN: 0167-6806, 1573-7217. DOI: 10/c22q7x. URL: http://link.springer.com/10.1007/s10549-011-1818-2 (visited on 03/06/2021).

[16] G. C. M. Siontis, I. Tzoulaki, K. C. Siontis, and J. P. A. Ioannidis. "Comparisons of established risk prediction models for cardiovascular disease: systematic review". en. In: *BMJ* 344.may24 1 (May 2012), e3318–e3318. ISSN: 1756-1833. DOI: 10/gf9hqx. URL: https://www.bmj.com/lookup/doi/10.1136/bmj.e3318 (visited on 03/06/2021).

[17] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. "Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures". en. In: *Epidemiology* 21.1 (Jan. 2010), pp. 128–138. ISSN: 1044-3983. DOI: 10/bj7bng. URL: https://insights.ovid.com/crossref?an=00001648-201001000-00022 (visited on 02/11/2020).

[18] S. van Dieren, J. W. J. Beulens, A. P. Kengne, L. M. Peelen, G. E. H. M. Rutten, M. Woodward, Y. T. van der Schouw, and K. G. M. Moons. "Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review". eng. In: *Heart (British Cardiac Society)* 98.5 (Mar. 2012), pp. 360–369. ISSN: 1468-201X. DOI: 10/fxzxxc.

[19] V. Mhasawade and R. Chunara. "Causal Multi-Level Fairness". In: *arXiv:2010.07343 [cs, stat]* (Oct. 2020). arXiv: 2010.07343. URL: http://arxiv.org/abs/2010.07343 (visited on 11/18/2020).

[20] Statistics Canada. *Canadian Community Health Survey - Annual Component (CCHS)*. eng. Jan. 2020. URL: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3226 (visited on 02/14/2020).

[21] Statistics Canada. *Canadian Community Health Survey – Annual component (CCHS)*. eng. Dec. 2019. URL: https://www.statcan.gc.ca/eng/survey/household/3226 (visited on 02/07/2020).

[22] Y. Béland. "Canadian community health survey—methodological overview". eng. In: *Health Reports* 13.3 (2002), pp. 9–14. ISSN: 0840-6529.

[23] Statistics Canada. *Vital Statistics - Death Database (CVSD)*. eng. Nov. 2019. URL: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3233 (visited on 02/14/2020).

[24] Statistics Canada. *Canadian Community Health Survey Data (2000 to 2011) Linked to the Canadian Vital Statistics Death Database (2000-2011)*. eng. Oct. 2018. URL: https://www.statcan.gc.ca/eng/rdc/cencchs-cvsd (visited on 10/06/2020).

[25] Statistics Canada. *Evaluation of the Health Statistics Program (2011/2012 to 2015/2016)*. eng. July 2020. URL: https://www.statcan.gc.ca/eng/about/er/hspfr (visited on 10/06/2020).

[26] Ontario Community Health Profiles Partnership. *Canadian Marginalization Index (CANMarg)*. 2020. URL: http://www.ontariohealthprofiles.ca/canmargCAN.php (visited on 02/10/2020).

[27] N. Ross, R. Wasfi, T. Herrman, and W. Gleckner. *Canadian Active Living Environments Database (Can-ALE) User Manual & Technical Document*. Tech. rep. Geo-Social Determinants of Health Research Group, Department of Geography, McGill University, 2018. URL: http://canue.ca/wp-content/uploads/2018/03/CanALE_UserGuide.pdf.

[28] Canadian Urban Environmental Health Research Consortium. *CANUE Metadata NDVI Landsat*. Aug. 2019. URL: https://canue.ca/wp-content/uploads/2019/09/CANUE-Browser-Metadata-NDVI-Landsat.pdf (visited on 02/14/2020).

[29] Canadian Urban Environmental Health Research Consortium. *CANUE Metadata Weather NRCAN*. May 2018. URL: https://canue.ca/wp-content/uploads/2018/11/CANUE-Metadata-Weather-NRCAN-Annual.pdf.

[30] Canadian Urban Environmental Health Research Consortium. *CANUE Metadata Annual Ozone*. Aug. 2019. URL: https://canue.ca/wp-content/uploads/2019/12/CANUE-Browser-Metadata-O3-CHG-Annual.pdf (visited on 02/14/2020).

[31] Canadian Urban Environmental Health Research Consortium. *CANUE Metadata SO2 OMI*. Aug. 2019. URL: https://canue.ca/wp-content/uploads/2019/09/CANUE-Browser-Metadata-SO2-OMI-Annual.pdf (visited on 02/14/2020).

[32] Canadian Urban Environmental Health Research Consortium. *CANUE Metadata PM2.5 DALb*. Aug. 2019. URL: https://canue.ca/wp-content/uploads/2019/09/CANUE-Browser-Metadata-PM25-DALb-Annual.pdf (visited on 02/14/2020).