


BMJ Open A case-control study on predicting population risk of suicide using health administrative data: a research protocol

JianLi Wang ,¹ Fatemeh Gholi Zadeh Kharrat,² Jean-François Pelletier,³ Louis Rochette,⁴ Eric Pelletier,⁴ Pascale Lévesque,⁴ Victoria Massamba,⁴ Camille Brousseau-Paradis,³ Mada Mohammed,¹ Geneviève Gariépy,^{5,6} Christian Gagné,² Alain Lesage⁷

To cite: Wang J, Gholi Zadeh Kharrat F, Pelletier J-F, *et al.* A case-control study on predicting population risk of suicide using health administrative data: a research protocol. *BMJ Open* 2023;**13**:e066423. doi:10.1136/bmjopen-2022-066423

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2022-066423>).

Received 06 July 2022

Accepted 12 February 2023

ABSTRACT

Introduction Suicide has a complex aetiology and is a result of the interaction among the risk and protective factors at the individual, healthcare system and population levels. Therefore, policy and decision makers and mental health service planners can play an important role in suicide prevention. Although a number of suicide risk predictive tools have been developed, these tools were designed to be used by clinicians for assessing individual risk of suicide. There have been no risk predictive models to be used by policy and decision makers for predicting population risk of suicide at the national, provincial and regional levels. This paper aimed to describe the rationale and methodology for developing risk predictive models for population risk of suicide.

Methods and analysis A case-control study design will be used to develop sex-specific risk predictive models for population risk of suicide, using statistical regression and machine learning techniques. Routinely collected health administrative data in Quebec, Canada, and community-level social deprivation and marginalisation data will be used. The developed models will be transformed into the models that can be readily used by policy and decision makers. Two rounds of qualitative interviews with end-users and other stakeholders were proposed to understand their views about the developed models and potential systematic, social and ethical issues for implementation; the first round of qualitative interviews has been completed. We included 9440 suicide cases (7234 males and 2206 females) and 661 780 controls for model development. Three hundred and forty-seven variables at individual, healthcare system and community levels have been identified and will be included in least absolute shrinkage and selection operator regression for feature selection.

Ethics and dissemination This study is approved by the Health Research Ethics Committee of Dalhousie University, Canada. This study takes an integrated knowledge translation approach, involving knowledge users from the beginning of the process.

INTRODUCTION

Suicide is a major international public health problem. Each year, over 4500 Canadians take their own life,¹ and more than 700 000 people die because of suicide worldwide,² imposing enormous impacts on families, communities

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ This study will use routinely collected health administrative data, which are readily accessible to policy and decision makers.
- ⇒ The candidate predictors include variables at individual, healthcare system and community levels, which reflect the complex aetiology of suicide.
- ⇒ The methodology of model development and validation needs to be improved.
- ⇒ Some individuals in the control group might have suicide behaviours, which could not be ascertained by health administrative data.
- ⇒ Important factors such as education, employment and income are not routinely collected by health administrative databases, which is a limitation of this study.

and societies. As such, suicide prevention has been a top priority of many countries.

Suicide has a complex aetiology and is a result of the interaction among the risk and protective factors at the individual, healthcare system and population levels.³⁻¹⁰ Therefore, policy and decision makers and mental health service planners can play an important role in suicide prevention. To facilitate suicide prevention planning, mechanisms should be in place that enable policy and decision makers to make informed decisions and mobilise resources to high-risk populations at the right places, before tragic events occur. This vision requires us to shift the paradigm from predicting individual risk to predicting population risk of suicide. However, the existing suicide risk assessment/predictive tools are not suitable for predicting population risk. Most of the existing risk assessment/risk predictive tools for suicide were designed to be used by clinicians; they were not designed for policy and decision makers.¹¹ Clinicians often use these tools to determine if individual patients are



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr JianLi Wang;
jianli.wang@dal.ca

at high risk of suicide presently or in short term (eg, next week). On the other hand, policy and decision makers are more concerned about the rate of suicide at the community level (eg, health regions, provinces/states) in the medium or long term (eg, in the next 5 or 10 years), driven partly by budgetary decisions that are often made on a yearly basis. Clinicians and policy/decision makers may have different emphases on risk predictive tools as well. For clinicians, an ideal suicide risk predictive tool should have high discriminative power (eg, a large C statistics), high sensitivity, specificity and positive predictive value. For policy and decision makers, a tool with excellent calibration (ie, how closely the predicted risk agrees with actual risk in the population) is more useful. To facilitate policy development in suicide prevention at the population level, risk predictive models specifically designed for policy and decision makers are needed.

Ideally, risk predictive models for population risk of suicide are based on large data from the target population. For example, Gradus *et al* developed sex-specific machine learning (ML) algorithms for suicide using data from eight Danish national health and social registries which cover more than 90% of the Danish population.¹² Kessler *et al*'s ML algorithms targeted US Army soldiers who were hospitalised.¹³ Accordingly, these risk predictive algorithms may potentially be used for forecasting the risk of suicide in Danish general population and in the US Army population, respectively. Furthermore, predictive models for population risk may use not only individual data, but also health system-level (eg, quality of mental healthcare, mental health budget) and community-level data (eg, unemployment rate and social deprivation levels in the community). For instance, Marks *et al* developed a predictive model for identifying counties at high risk of overdose mortality, which included county-level education, poverty rate, unemployment rate, overdose gravity and other county-level indicators, among the 3106 counties in the USA.¹⁴ Given the complex aetiology of suicide, predicting population risk of suicide may benefit greatly from the integration of data at the individual, health system and community levels.

We undertook a project to develop and validate sex-specific risk predictive models to be used by policy and decision makers to forecast population risk of suicide at the health region level, using routinely collected health administrative data, and to identify the barriers and facilitators to implementation and explore the ethical and privacy issues of the prediction program. In this manuscript, we aimed to describe the methodology of the project, to inform methodological discussions and suicide prevention strategies.

METHODS

This project encompasses the components of quantitative and qualitative investigations and an integrated knowledge translation (IKT). IKT is a model of research co-production, whereby knowledge users are integrated

throughout the research process and who can use the research recommendations in practice or policy.¹⁵ IKT approaches are used to improve the relevance and impact of research. The quantitative research involved developing and validating risk prediction models for suicide using advanced ML and visualisation methods. The qualitative research is to understand the potential implementation, social, ethics and legal issues associated with the risk prediction program. In line with IKT principles, we involved policy and decision makers at the provincial and national levels, and people with lived experience of suicidality from the beginning of the project. The methodology of each component is described below.

Model development and validation

Target population

The general population residing in the province of Quebec, Canada. The province had a population of over 8.6 million people in 2021, and about 95% of the population reported being able to conduct a conversation in French. In Quebec, health services are planned and delivered through 18 health regions, 22 integrated health and social services centres and 166 Centres locaux de santé Communautaire. Budgetary decisions are made at the levels of province and health regions/integrated health and social services centres.

Data sources

We will develop the prediction tools by linking the suicide database, the Ministry of Health and Social Services public financial reports (Contour financier-Publications du ministère de la Santé et des Services sociaux (gouv. qc.ca)), which include the five health administrative databases below, and the Canadian Urban Environmental Health Research (CANUE) data. The suicide database gathers individual-level data annually based on residents' health insurance number from five administrative databases: the vital statistics death database, the physician claims database, the hospital discharge database, the Insured Person Registration File and the public drug plan. The data of these databases (eg, billing and service procedures codes, service dates) are routinely submitted by clinics and hospitals for billing and administration purposes; no self-reported data were collected from patients. These databases cover up to 98% of the population in Quebec and contain data for over 20 000 death by suicide cases that occurred since 1996. Death by suicide cases were those ascertained by Quebec's Coroner office after investigation. The decision is registered in the Quebec vital statistics database. The latter is linked with other health administrative databases of the Quebec Integrated Chronic Disease Surveillance System (QICDSS) managed by the Quebec's Public Health Agency.⁵ With the suicide database and other linkable Ministry financial databases, individual (eg, sex, age), programme (eg, hospitalisation, emergency department visits) and system (eg, mental health and addiction budgets) level indicators can be identified.⁵

CANUE is a Canadian consortium aiming to build a unique repository of standardised metrics of urban, suburban and rural characteristics, as well as the tools used to produce them (www.canue.ca). The CANUE data contain indicators for unemployment, social deprivation, access to health services and built environment at the community level, and can be linked with health administrative data by postal codes. The CANUE is open and free for research projects. The data linkage was performed at the Quebec Institute of Public Health (INSPQ) where the suicide data are kept. Linking the databases provides an unprecedented sample size and the capability of examining individual, neighbourhood, programmatic and systemic indicators of population suicide risk.

Because this study used existing de-identified health administrative data, informed consent from individual patients was waived. This study was approved by the Research Ethics Board of Dalhousie University.

Study design

Because the base rate of suicide in the population is low, we proposed using a case–control study design to develop sex-specific suicide risk predictive algorithms, using both logistic regression modelling and ML techniques. We selected all death by suicide cases that occurred from 1 January 2002 to 31 December 2010.¹⁶ The control group was a 1% random sample of living individuals in each year between 1 January 2002 and 31 December 2010 from the Quebec physician claim database. Controls are not allowed to be selected more than once across years. None of those in the control group died of suicide during this period. The cases and controls were not matched to allow for maximum variability in predictors.

Predictors

Individual, programmatic, systemic and community factors (see online supplemental appendix 1) that happened 5 years prior to the suicide events will be used as candidate predictors to develop the risk predictive algorithms. For example, we extracted the data about the diagnosis of major depression (an individual-level factor) in the past 6, 12, 24, 48 and 60 months, as five separate candidate predictors. Similarly, we extracted mental health and addiction budget of each health region (a systemic-level factor) in the past 5 years as candidate predictors. The QICDSS¹⁷ provided all the variables drawn from health administrative databases. It covers 98% of the Quebec's population since 1996. The security and continuous quality and maintenance are the responsibility of the INSPQ. Information is for administrative (ie, age, hospital or outpatient contact dates) and clinician reporting (ie, diagnoses) purposes. Validation of QICDSS physical diagnoses has been achieved by chart reviews¹⁷ and by outcomes for QICDSS psychiatric diagnoses.^{18 19} The QICDSS has been exploited over the past decade by a network of INSPQ officers and academic researchers, many are coauthors of publications on the characteristics of patients receiving rare psychiatric interventions,²⁰ and

on personality disorders, schizophrenia and substance use disorders in relation to mortality, including suicide.^{21 22} The quality of the data is also reflected by the minimal missing data associated with the variables, which range from 0.87% and 4.12% of the variables in the databases.

The initial selection of candidate predictors is determined by content knowledge (ie, known relationships between suicide or suicide behaviours and individual and local area-level variables), feasibility of routine data collection, clinical utility and policy relevance through team meetings. Therefore, the predetermination of candidate predictors was a joint effort between the team members, collaborators, health policy and decision makers and other stakeholders, with the expertise of clinical psychiatry, psychiatric epidemiology, mental health services research, health administrative data, computer science and mental health policy.

For the objective of this study, we will use both statistical (eg, logistic regression modelling) and ML approaches to develop the risk prediction models so that we may compare which approach performs better in predicting population suicide risk and is more feasible to implement. ML can produce complex estimations by searching data for relevant pieces of information and their complex interactions. Therefore, ML is best suited to tackle the combined challenges of high-dimensional data analysis associated with risk prediction for suicide. Some predictors that may change over time (eg, diagnoses, medications, service use, etc) will be dummy-coded to create time-varying predictors (ie, intervals of 0–3, 0–6, 0–12, 0–24, 0–36, 0–48 and 0–60 months before the first day of the suicide month). Because we included all suicide cases and a sample of controls, the proportion of suicide in the sample is different from that in the general population. This is a limitation of case–control study design which produces a biased sample because the proportion of cases in the sample is not the same as the population of interest.^{23 24} One method for addressing this limitation when developing predictive models using case–control data is weighting.^{24–27} Therefore, in logistic regression modelling, sampling weights (inverse probability of being selected) were assigned to the controls, while the weight of 1 was assigned to the cases, to ensure the models are applicable to the whole population.

Model development: ML

ML is a part of artificial intelligence that aims to construct systems that automatically improve through experience using advanced statistical and probabilistic techniques. ML has provided significant benefits to a range of fields. Recent research has shown a range of advantages of ML that can assist in detecting, diagnosing, predicting suicide and treating mental health problems.^{28 29} ML methods are divided into categories, that is, supervised, semisupervised, unsupervised and reinforcement.

Imbalanced classes are a common problem in ML classification, where each class has a disproportionate ratio of observations. To predict the population risk of suicide,

dataset will be imbalanced because of rare cases of suicide as compared with a control group. To address the imbalanced dataset, we will oversample the minority class. We will 'artificially' duplicate samples from the minority class to oversample the minority class to correct imbalanced datasets, even though doing so does not provide the model with any new data. In the literature, this method was known as the Synthetic Minority Over-sampling Technique. Then, we will develop supervised learning models such as logistic regression, random forest, XGBoost and multilayer perceptron with an optimised model architecture. These models' predictive capacity will be assessed by generating the receiver operating characteristic curves calculating its area under the curve and various operating characteristics, including sensitivity, specificity and positive predictive value for a variety of thresholds.

Interpretability is essential when we deal with health-care data. It is significant because it is necessary to understand the causality of learnt representations for decision support also helps to assess whether the model is considering the right features while making a specific prediction. Feature-based model explainability technique, such as Shapley Additive Explanations (SHAP), was derived from game theory; each player decides to contribute to a coalition of players to produce a total value that will be superior to the sum of their individual values. SHAP relies on the Shapley value of both local and global explanations. Shapley's values are model-agnostic, and the marginal contribution of each feature can be calculated by using the input data and the predictions.^{30 31} SHAP will use the global explanation of how much the input features contribute to a model's output.

Model development: logistic regression

As the first step of model development, we will include all preselected variables in penalised least absolute shrinkage and selection operator (LASSO) regression. The LASSO penalisation factor selects important predictors by shrinking coefficients for weaker predictors toward zero, excluding predictors with estimated zero coefficients from the final sparse prediction model. We will perform a correlation analysis among variables selected by the LASSO regression, and identify variables that are strongly correlated (eg, $\gamma \geq 0.60$). Correlated variables will be discussed by team members, and the variables that have better policy implication and clinical utility will be kept and become the candidate predictors for model development.

We will use logistic regression to develop the sex-specific statistical models. After LASSO, there may still be a large number of candidate predictors. Backward selection method will be used to eliminate uninformative variables and to identify the model with the best calibration and discrimination. The decisions of model selection will be initially based on the changes in the values of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).³² Since BIC penalises for the complexity of the model more than AIC, selection with BIC will generally lead to smaller models than

selection with AIC.³² Once a model is developed, prediction accuracy will be assessed by the discrimination and calibration of the model. Discrimination is the ability of a prediction model to separate those who experienced the outcome events from those who did not. We will quantify this by calculating the C statistic, analogous to the area under a receiver operating characteristic curve. Calibration measures how closely predicted outcomes agree with actual outcomes. For this, we will use D'Agostino's version of the Hosmer-Lemeshow X^2 statistic. Discrimination and calibration compete with each other. Given that the program will be used to forecast population risk of suicide, we will prioritise calibration over discrimination. Stakeholders from different perspectives and scientific backgrounds will meet to determine the content and performance of the risk prediction models developed by statistical and ML techniques, the appropriate formats of data visualisation that are acceptable to policy and decision makers, and the feasibility of implementation, which will in turn inform the revision of the models.

The second step of the model development is to estimate the synthetic rates, consisting of two stages. First, for each predictor, the proportions of individuals within each category of that predictor in the initial modelling will be computed, separately by regions. For instance, if hospitalisation due to suicide attempt in the past 5 years is a predictor in the model, the proportion of individuals with this attribute in a specific health region is calculated. If age is a continuous variable in the model, the mean age of the population in a health region is estimated. A syntax program will then be prepared to apply the regression coefficients to the corresponding proportions and means in the dataset, and to calculate the logit estimates for each of health regions. The resulting logit values for each of the health region will then be converted into probabilities, giving the estimated risk of suicide in the health region. The region's population counts from Statistics Canada Census data or the provincial health administrative database multiplied by the estimated risk will yield the estimated number of suicide in this health region.

The fitted logistic regression model described above estimates the proportion of suicide in the population at a given moment of time as a function of its risk factors in the past. This model is fundamentally aetiological, where the natural reference point is the moment of the outcome's occurrence, corresponding to the zero time on the aetiological time scale. However, assessment of population risk of suicide over a particular span of time in the future involves a prognostic outlook, where the natural reference point is the time of prognostication, corresponding to the zero time on the prognostic time scale. Predictive models for individual risk are often developed using a cohort/closed study population and express the risk of future occurrence of the outcome as a function of current risk factors, and involve consideration of the values of the risk factors at issue at the prognostic time zero only. On the other hand, population risk models are applied in the context of a dynamic/open population and the estimated risk is a function of risk factors not only at the

prognostic time zero but also throughout the time span at issue. For example, the risk of suicide in the next 5 years in a health region may not only depend on the proportions of people with major depression and of hospitalisation due to suicide attempt in the past, but also on whether there will be a reduction or increase in these parameters over the next 5 years, if so in which year. Thus, the population risk of suicide may be projected using the developed model to each future year over a predefined time interval. The cumulative incidence of suicide ($CI_{0 \text{ to } t}$) from time $T=0$ to $T=t$ can be estimated as a function of time-specific and profile-specific risk operating over that time interval²⁷:

$$CI_{0 \text{ to } t} = 1 - \exp \left[- \int_0^t (ID_u) du \right]$$

The estimated cumulative risk represents the estimated risk of suicide of a health region over the time period at issue conditionally on the health region's risk profile.

Validation

For model validation, we will use the suicide data from 1 January 2011 to 31 December 2019. We will first calculate the yearly, 5-year and 10-year incidence of suicide death at the provincial and health regional levels in males and females (ie, observed risk). We will apply the developed synthetic models in the validation data to estimate the yearly, 5-year and 10-year incidence of suicide death at the provincial and health regional levels in males and females (ie, predicted risk). We will visually compare and calculate the differences between the predicted and observed risks; smaller differences indicate better calibration with the data and model accuracy. We will use four indicators for assessing model performance: mean average error (MAE), root mean square error (RMSE), Spearman's r and proportion of correct identification of high-risk regions.¹⁴ The MAE is the average magnitude of the difference between the predicted and observed suicide death rate for each health region. The RMSE is the square root of the average magnitude of the difference squared, therefore is similar to MAE but penalises prediction errors with greater magnitude. More accurate predictions will result in smaller MAE and RMSE. Spearman's r compares the predicted ranking of health regions by suicide death rate compared with the actual observed rankings; results closer to 1 indicate that the model was more effective at rank-ordering regions based on suicide death rate. To assess the extent to which high-risk regions are correctly identified, we will first disaggregate the predicted and observed suicide rates into quartile groups and categorised all health regions into their corresponding quartiles for both predicted and observed suicide rates. The proportion of health regions observed in the top quartile of observed suicide death rates that were rightly predicted to be in the top quartile will be calculated.

Qualitative study

The objective of the qualitative study is to investigate the end-users' views about predicting population risk of suicide, and the potential social, legal, ethical, and privacy issues and mitigation strategies for implementing

such a predictive system. Using snowballing techniques, we have invited policy and decision makers at the federal and provincial levels, mental health professionals, individuals who have extensive experience in working with policy and decision makers and who have expertise in suicide prevention, social and health policy, as well as health administrative data, people with lived experience and advocates for families bereaved by suicide. The qualitative study consists of two rounds of interviews. The first round of interviews was carried out after the general team meeting held in July 2021, at which the study design was finalised. The second round of interviews will be organised once the predictive models are developed. The first round of interviews was held through zoom meetings, and followed a series of semistructured interview questions related to the objectives (see online supplemental file 1). Qualitative data collected during the focus groups and qualitative interviews are audio recorded, transcribed and analysed with the support of QDA Miner (Provalis).³³ The second round of interviews will be conducted once the prototype models are developed and presented at the second general team meeting which is to be held in late 2022. We will perform an inductive thematic analysis of the focus group and individual interview material, which will be fed by answers to the open questions regarding potential (1) perceptions about the developed prediction models, (2) social issues, (3) legal issues, (4) ethical and privacy issues, and (5) mitigation strategies for implementing such a system. Transcripts will be coded in order to demarcate segments within each of them. We will look for words or short phrases that demonstrate how the associated data segments inform our research objectives. Detailed results from the qualitative analysis of this material will be presented in a separate paper.

Patient and public involvement

Engagement with relevant stakeholders (eg, policy/decision makers and people with lived experience) through IKT is critical for developing equitable risk predictive algorithms and for maximising the potential for future implementation. For this project, we have identified and engaged policy/decision makers from the Public Health Agency of Canada and from the INSPQ, as well as eight people with lived experience. The representatives of INSPQ (EP, PL, VM, LR) were involved in study conceptualisation and grant application. PL has been facilitating data extraction and participated in the biweekly team meetings. As described above, we have engaged people with lived experience through the qualitative interviews. The next round of qualitative interviews will be held after the prototype of the risk predictive models is developed to have a better understanding about privacy, ethics and implementation issues.

ETHICS AND DISSEMINATION

This study will use routinely collected health administrative data. The analysis of secondary de-identified data at the INSPQ where the data are kept will not incur physical

and psychological harms. The results of the study will be vetted by analysts at the INSPQ to ensure no privacy and confidentiality will be breached. The data used for this study will be kept at INSPQ for 15 years. The results will be presented in peer-reviewed journals, at academic conferences and shared with knowledge users who were engaged from the beginning.

Through this study, we aimed to develop risk prediction models to be used by policy and decision makers to forecast population risk of suicide at the provincial and health region levels, using routinely collected health administrative data and other publicly available area-level data. For example, policy and decision makers may use the models to project the proportion and number of suicide deaths in specific health regions/communities over the next 5 years, and decide how resources and community-level interventions may be mobilised to the high-risk regions/communities. Furthermore, the models can inform policy and decision makers about the potential impacts of these community-level interventions on suicide prevention. The potential utility of such predictive tools has been attested by the active involvement by the policy and decision makers at the federal and provincial levels and people with lived experience. Nevertheless, predicting population risk of suicide is new and has not been well studied. There are a number of methodological and implementation challenges to be addressed.

Routinely collected health administrative data and population health survey data represent a unique opportunity for population health projection because it covers a majority of the general population in catchment areas, and the data can be readily accessed by policy and decision makers. Many risk predictive models have been developed for physical and mental health problems in the general population. For example, individual data from population health surveys and health administrative databases have been used to develop risk predictive models for diabetes,³⁴ heart disease³⁵ and major depression.^{36 37} These models may be used to identify high-risk individuals in the community; they can also be used to forecast the population risk in the future. However, few models have integrated individual, healthcare system and community-level predictors in the same model. In this study, we proposed including data from these different levels in model development, and converting the models into synthetic estimation models. There may be different approaches for integrating data from different levels for population risk prediction. Future studies are needed to explore the best method for data integration.

The performance of a risk predictive model is commonly assessed by indicators of model discrimination and calibration.³⁸ Whereas model discrimination is critical for individual risk predictive models, policy and decision makers' focus is on the whole population rather than individuals. Therefore, model calibration plays a more important role in the performance of a population risk model. We proposed four indicators for assessing model performance. However, it is not clear how much error (the difference between predicted

and observed risks) policy and decision makers may tolerate for population risk prediction, how they perceive the importance of model discrimination and whether other indicators exist for assessing population risk prediction models. We will explore these aspects through our qualitative study, and also encourage others to consider these in future studies. Similarly, we welcome discussions and debates about the methods for validating population risk predictive models. An individual risk predictive model is often developed using longitudinal cohort/closed population data and validated in a different but related cohort/closed population. This poses challenges for population risk predictive models because the population in a community/health region is open and dynamic. Appropriate methods for model validation and acceptability need to be developed and agreed by the research community and policy and decision makers.

This study relied on routinely collected health administrative data for model development and validation, rather than collecting primary data. Therefore, we have little information about suicide behaviours among the individuals in the control group, which are strongly associated with suicide deaths. In the model development, we included hospitalisation and emergency department visits due to suicide attempt, which may reduce the bias related to the lack of information about suicide behaviours. Nevertheless, this is a limitation of routinely collected health administrative data.

Despite the challenges for developing population risk predictive model for suicide, research is urgently needed to address this important population health issue. This study represents one of the early steps in building such risk predictive models and methodology development, as part of the collective efforts for moving the field forward.

Author affiliations

¹Department of Community Health and Epidemiology, Dalhousie University, Halifax, Nova Scotia, Canada

²Department of Electrical Engineering and Computer Engineering, Laval University, Quebec, Quebec, Canada

³Department of Psychiatry, University of Montreal, Montreal, Québec, Canada

⁴Institut national de sante publique du Quebec (INSPQ), Quebec City, Quebec, Canada

⁵Public Health Agency of Canada, Ottawa, Ontario, Canada

⁶Department of Social and Preventive Medicine, University of Montreal, Montreal, Québec, Canada

⁷Institut universitaire en sante mentale de Montreal, Montreal, Québec, Canada

Contributors JLW drafted the manuscript. JLW, FGZK, J-FP, LR, EP, PL, GG, CG and AL were involved in study design, conceptualisation and funding application. JLW, FGZK, J-FP, LR, EP, PL, VM, CB-P, MM, GG, CG and AL were involved in manuscript review, discussion, revision and final approval.

Funding This study is supported by a New Frontiers for Research Funds grant (2019-00471) from Tri-Agency Institutional Programs Secretariat, Government of Canada and by a Tier I Canada Research Chair award to JLW.

Disclaimer The funders play no role in the design and operation of this study.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and

responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

JianLi Wang <http://orcid.org/0000-0002-1329-914X>

REFERENCES

- 1 Statistics Canada. Deaths and age-specific mortality rates, by selected grouped causes. 2019. Available: <https://www150.statcan.gc.ca/t1/tb11/en/tv.action?pid=1310039201>
- 2 World Health Organization. Suicide. Available: <https://www.who.int/news-room/fact-sheets/detail/suicide> [Accessed 15 May 2022].
- 3 Mental Health Commission of Canada. Research on suicide and its prevention: what the current evidence reveals AND topics for future Research. Ottawa, Canada; 2018. Available: www.mentalhealthcommission.ca
- 4 Tondo L, Albert MJ, Baldessarini RJ, et al. Suicide rates in relation to health care access in the United States: an ecological study. *J Clin Psychiatry* 2006;67:517–23.
- 5 Thibodeau L, Rahme E, Lachaud J, et al. Individual, programmatic and systemic indicators of the quality of mental health care using a large health administrative database: an avenue for preventing suicide mortality. *Health Promot Chronic Dis Prev Can* 2018;38:295–304.
- 6 Mortier P, Cuijpers P, Kiekens G, et al. The prevalence of suicidal thoughts and behaviours among college students: a meta-analysis. *Psychol Med* 2018;48:554–65.
- 7 Nock MK, Borges G, Bromet EJ, et al. Suicide and suicidal behavior. *Epidemiol Rev* 2008;30:133–54.
- 8 Goldman-Mellor SJ, Caspi A, Harrington H, et al. Suicide attempt in young people: a signal for long-term health care and social needs. *JAMA Psychiatry* 2014;71:119–27.
- 9 Crawford MJ, Nur U, McKenzie K, et al. Suicidal ideation and suicide attempts among ethnic minority groups in England: results of a national household survey. *Psychol Med* 2005;35:1369–77.
- 10 Bernal M, Haro JM, Bernert S, et al. Risk factors for suicidality in Europe: results from the ESEMED study. *J Affect Disord* 2007;101:27–34.
- 11 Bolton JM, Gunnell D, Turecki G. Suicide risk assessment and intervention in people with mental illness. *BMJ* 2015;351:h4978.
- 12 Gradus JL, Rosellini AJ, Horváth-Puhó E, et al. Prediction of sex-specific suicide risk using machine learning and single-payer health care registry data from Denmark. *JAMA Psychiatry* 2020;77:25–34.
- 13 Kessler RC, Warner CH, Ivany C, et al. Predicting suicides after psychiatric hospitalization in US army soldiers: the army study to assess risk and resilience in servicemembers (army Stars). *JAMA Psychiatry* 2015;72:49–57.
- 14 Marks C, Abramovitz D, Donnelly CA, et al. Identifying counties at risk of high overdose mortality burden during the emerging fentanyl epidemic in the USA: a predictive statistical modelling study. *Lancet Public Health* 2021;6:e720–8.
- 15 Boland L, Kothari A, McCutcheon C, et al. Building an integrated knowledge translation (IKT) evidence base: colloquium proceedings and research direction. *Health Res Policy Syst* 2020;18:8.
- 16 Vasilopoulos HM, Lesage A, Latimer E, et al. Implementing suicide prevention programs: costs and potential life years saved in Canada. *J Ment Health Policy Econ* 2015;18:147–55.
- 17 Blais C, Jean S, Sirois C, et al. Quebec integrated chronic disease surveillance system (QICDSS), an innovative approach. *Chronic Dis Inj Can* 2014;34:226–35.
- 18 Lesage A, Rochette L, Émond V, et al. A surveillance system to monitor excess mortality of people with mental illness in Canada. *Can J Psychiatry* 2015;60:571–9.
- 19 Diallo FB, Pelletier É, Vasilopoulos H-M, et al. Morbidities and mortality of diagnosed attention deficit hyperactivity disorder (ADHD) over the youth lifespan: a population-based retrospective cohort study. *Int J Methods Psychiatr Res* 2022;31:e1903.
- 20 Lafrenière S, Gholi-Zadeh-Kharrat F, Sirois C, et al. The 5-year longitudinal diagnostic profile and health services utilization of patients treated with electroconvulsive therapy in Quebec: a population-based study. *Soc Psychiatry Psychiatr Epidemiol* 2022.
- 21 Huynh C, Kisely S, Rochette L, et al. Using administrative health data to estimate prevalence and mortality rates of alcohol and other substance-related disorders for surveillance purposes. *Drug Alcohol Rev* 2021;40:662–72.
- 22 Cailhol L, Pelletier É, Rochette L, et al. Prevalence, mortality, and health care use among patients with cluster B personality disorders clinically diagnosed in Quebec: a provincial cohort study, 2001–2012. *Can J Psychiatry* 2017;62:336–42.
- 23 Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959;22:719–48.
- 24 Rose S, van der Laan MJ. A note on risk prediction for case-control studies [Preprint]. 2008.
- 25 van der Laan MJ. Estimation based on case-control designs with known prevalence probability. *Int J Biostat* 2008;4:Article 17.
- 26 Haaf KT, Steyerberg EW. Methods for individualized assessment of absolute risk in case-control studies should be weighted carefully. *Eur J Epidemiol* 2016;31:1067–8.
- 27 Karp I, Sylvestre MP, Abrahamowicz M, et al. Bridging the etiologic and prognostic outlooks in individualized assessment of absolute risk of an illness: application in lung cancer. *Eur J Epidemiol* 2016;31:1091–9.
- 28 Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. *Psychol Med* 2019;49:1426–48.
- 29 McHugh CM, Large MM. Can machine-learning methods really help predict suicide? *Curr Opin Psychiatry* 2020;33:369–74.
- 30 Bowen D, Ungar L. Generalized SHAP: generating multiple types of explanations in machine learning. *ArXiv* 2020.
- 31 Sundararajan M, Najmi A. The many shapley values for model explanation. *ArXiv* 2019.
- 32 Vrieze SI. Model selection and psychological theory: a discussion of the differences between the akaike information criterion (AIC) and the Bayesian information criterion (Bic). *Psychol Methods* 2012;17:228–43.
- 33 Rivas C, Tkacz D, Antao L, et al. Automated analysis of free-text comments and dashboard representations in patient experience surveys: a multimethod co-design study. *Health Serv Deliv Res* 2019;7:1–160.
- 34 Rosella LC, Manuel DG, Burchill C, et al. A population-based risk algorithm for the development of diabetes: development and validation of the diabetes population risk tool (dport). *J Epidemiol Community Health* 2011;65:613–20.
- 35 Manuel DG, Tuna M, Bennett C, et al. Development and validation of a cardiovascular disease risk-prediction model using population health surveys: the cardiovascular disease population risk tool (CVDPORT). *CMAJ* 2018;190:E871–82.
- 36 Wang JL, Manuel D, Williams J, et al. Development and validation of prediction algorithms for major depressive episode in the general population. *J Affect Disord* 2013;151:39–45.
- 37 Wang J, Sareen J, Patten S, et al. A prediction algorithm for first onset of major depression in the general population: development and validation. *J Epidemiol Community Health* 2014;68:418–24.
- 38 Steyerberg EW. *Clinical prediction models. A practical approach to development, validation, and updating*. Springer, 2009.