PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	A case-control study on predicting population risk of suicide using health administrative data: A research protocol
AUTHORS	Wang, JianLi (proxy) (contact); Gholi Zadeh Kharrat, Fatemeh; Pelletier, Jean-François; Rochette, Louis; Pelletier, Eric; Lévesque, Pascale; Massamba, Victoria; Brousseau-Paradis, Camille; Mohammed, Mada; Gariépy, Geneviève; Gagné, Christian; Lesage, Alain

VERSION 1 – REVIEW

REVIEWER	Sequeira, Lydia University of Toronto Institute of Health Policy Management and Evaluation
REVIEW RETURNED	29-Jul-2022

GENERAL COMMENTS	Thank you for the opportunity to review this manuscript. The authors detail a protocol for a nested case-control study design for developing risk predictive models for population risk of suicide. The use of multiple data bases is a true strength of this project, aiming to use a multitude of predictors for this outcome. Below are a few comments for consideration to help strengthen the manuscript/ protocol: • Within study design section (page 5), it would be helpful to clarify that the suicide cases were “death by suicide”. Additionally, did the controls have any previous suicidal behaviour (i.e. ideation, attempts) – this level of detail also helps clarify the data used for model generation. • Within the predictors section (page 5), it would be helpful to add some detail about data quality for different types of predictors that you are using. Are there large amounts of missing data? Is any routine quality check done across databases? Additionally, some details about whether these variables are all clinician-reported, or if some are gathered through patient self-report would be helpful. • If possible, could you add an appendix with your qualitative interview guide to provide the reader with additional context? • A comment or two about anticipated use cases of how policy makers and/or decision makers are hoping to use these models would be helpful – is it to predict yearly suicide rates and have initiatives/ campaigns accordingly? Is there the ability to make regional predictions (i.e. more targeted suicide prevention initiatives)?

REVIEWER	Yin, Honglei Southern Medical University
REVIEW RETURNED	15-Sep-2022

GENERAL COMMENTS	In the manuscript: "A case-control study on predicting population risk of suicide using health administrative data: A research protocol" the authors describe the rationale and methodology for developing risk predictive models for population risk of suicide. I believe this is an interesting study, with comprehensive theoretical analysis, and therefore I recommend publication after revision. I have a few minor comments for the authors to consider: There was some ambiguity in the definition of suicide. The author did not describe the difference between non-Suicidal Self-Injury and suicidal Behaviour, single and Multiple Suicide Attempters, or suicide Attempters and suicide Ideators, whicn will have an enormous influence in the results of models. "Individual, programmatic, systemic and community factors (see Appendix I) measured five years prior to the suicide events were used to develop the risk predictive algorithms. " The sentence is very unclear. A detailed explanation is needed on the definition of suicide events, and how they can be measured. The educational level, income level, employment status and marital status were also associated with the risk of suicide.The risk predictive models for population risk of suicide appear not appropriate without considering these factors. The author used a nested case-control study, but the method seems to be inappropriate for this study. In the nested case-control study, cases of a disease that occur in a defined cohort are identified and, for each, a specified number of matched controls is selected from among those in the cohort who have not developed the disease by the time of disease occurrence in the case. In this study, the author used routinely collected health administrative data, without follow-up. Suicide cases and control group were not selected from a same group of samples matching certain conditions. Thus, this study seems to be more like a retrospective study rather than a nested case-control study.

REVIEWER	Mortier, Philippe Hospital del Mar Institute for Medical Research, Health Services Research Group
REVIEW RETURNED	02-Oct-2022

GENERAL COMMENTS	This is a protocol paper that presents a methodology to enable the forecasting of suicide risk at the health region level, as well as the simulation of healthcare policy changes on suicide risk at the health region level. Key elements are the use of machine-learning techniques, a nested-case control study design, and the use of various sources of data (suicide data, Health and Social Services public financial reports, and Canadian Urban Environmental Health Research data). The study protocol’s general objective is interesting. However, I am afraid that the protocol paper in its current form lacks methodological clarity on various critical points, further described below. This makes it difficult to evaluate the scientific integrity of the proposed methodology. Also, it seems that data collection is finished and the database fully constructed, which would make this paper not eligible for publication in the journal. This needs to be checked with the editor, given the fact that this is a retrospective register-based study.

	Critical points: A first critical point is that the study design is not clearly described. A nested case-control study is a case-control study nested within a cohort (or within a case-cohort). Once the cohort is clearly defined, a number of controls are selected for each individual case, or for the number of cases occurring in a certain time-unit (day, week, month, year), using a certain case-control ratio. The cohort is not clearly described in the paper, and is not clear which administrative information was used to create the cohort. Information is needed on mortality, birth, immigration and emigration in / out of the population in order to create the retrospective cohort. Which information was used? It would be interesting to know why was the period 2002-2010 was chosen. The authors state that (page 6, line 20) “the control group was a 1% random sample of living individuals in Quebec each year”. That does not seem to be conform a nested case-control study. Were controls allowed to be selected more than once across years? Were future cases eligible? Critically, why sample 1% of controls each year? Why not a number proportional to the number of cases in each year? A second critical point is the clarity and rationale for the proposed model development. Both logistic regression and machine learning are proposed, and it seems that they serve different objectives. Is that correct? It is not clear (at least, not early on in the paper) which objectives exactly are reached by which techniques, and why different techniques are needed. Machine learning is proposed to be used to enable simulations of policy changes on suicide risk rates, while logistic regression seems to be proposed to be used to forecast future suicide risk based on past data. Why this differentiation in techniques according to different objectives? A third critical point is that the use of sampling weights is not clear. Why are these sampling weights needed exactly? How are they calculated? Critically, in which analysis (and how) are they applied? Also, it seems to me that class imbalance is an equally important issue but this is nowhere addressed. A fourth critical point refers to what the authors state in the paragraph on Page 8 (paragraph starting at lines 25-26). I do not clearly see the difference between individual-level models and population-level models that the authors introduce here. They state “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 hospitalization 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.” I think that this is not unique to population-level models. An individual-level (clinical) model can be developed using data available up to time x (e.g., a certain healthcare visit) and the outcome can be situated at time y, with y being close (days-weeks) to far (months-years) from time x. Any changes in the risk factors of the patient between x and y will not be taken into account by the model. Major issues: Abstract: the authors introduce the idea of population risk versus individual risk, but mention nowhere that the objective of the study is to “forecast population risk of suicide at the health region level”, This should be made clear.

	Introduction (page 4, line 6): the authors state that “Suicide has a complex etiology and is a result of the interaction among the risk and protective factors at the individual, healthcare system, and population levels.(3–5)”. The cited papers do not represent key papers (e.g., systematic reviews, meta-analyses) on the various types of risk and protective factors for suicide, while many of those key papers are available in the literature. Methods - Model development and validation - Data sources (page 5, line 40): while the suicide database and CANUE data are fairly well described, there is no description of the Ministry of Health and Social Services (MSSS) public financial reports. A clear description of all used data should be provided. Methods - Model development and validation – Predictors (page 6, line 28): The authors state that “The selection of candidate predictors are determined by content knowledge (i.e., known relationships between suicide or suicide behaviors and individual and local area level quality of health care), feasibility of routine data collection, clinical utility and policy relevance”. No literature is cited to back up this statement. Methods - Model development and validation - Machine learning (page 7, line 7): The authors state that “From that, we will make simulations of changes coming from policies by modifying the population composition for reflecting the effect of policies change (e.g., mental health diagnoses, socio-economic factors, health system resources allocation) and evaluate their effect on the suicide rates predicted, comparing these with rates obtained with the current population and population modified differently.” This was not stated as an objective previously. It should be made more clearly how this will be done, and why machine learning is needed for this. Model development – logistic regression: Why is a correlation analysis among variables selected by the LASSO regression needed? LASSO will eliminate them. See the paper by Cox et al. (2020) for a discussion of this topic. Methods - Model development and validation (page 7, line 41): The authors state that “That accounted for clusterings at the health regions”. How exactly will this be done? Methods - Model development and validation (page 7, line 41): the authors state that “Backward selection method will be used to identify the model with the best calibration and discrimination”. Backward selection, as I know it, is a method to select variables to include in a model. The use of a “backward selection method” needs to be clarified in the context of model selection. Especially since LASSO was earlier described as a predictor selection method. Methods - Model development and validation (page 8, line 11): The authors state that “First, for each predictor, the proportions of individuals within each category of that predictor in the initial modeling will be computed, separately by regions. For instance, if hospitalization 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.” This seems like a rather elaborate way to explain that aggregate values by health region of predictor variables will be calculated. Is this correct? Minor issues: Abstract: the author mention “synthetic estimation models”. It is not clear what is meant by “synthetic”.

	Methods - Model development and validation (page 4, line 52): the authors state “and the indicators at each level may be classified into the broad categories of input and process”. It is not clear what is meant by “input and process”. References Cox, C. R., Moscardini, E. H., Cohen, A. S., Tucker, R. P. (2020). Machine learning for suicidology: A practical review of exploratory and hypothesis-driven approaches. Clinical Psychology Review, 82(October), 101940. https://doi.org/10.1016/j.cpr.2020.101940

VERSION 1 – AUTHOR RESPONSE

Responses to reviewer# 1’s comments

1. *Within study design section (page 5), it would be helpful to clarify that the suicide cases were “death by suicide”. Additionally, did the controls have any previous suicidal behaviour (i.e. ideation, attempts) – this level of detail also helps clarify the data used for model generation.*

The definition of suicide has been clarified in the Methods (last paragraph, page 4).

Because we used health administrative data which is mainly for administration and reporting purposes, no data about previous suicidal behaviors are routinely collected. However, we can extract data about hospitalization and emergency department visits due to suicide attempts or self-injuries, and include the variables as potential predictors.

2. *Within the predictors section (page 5), it would be helpful to add some detail about data quality for different types of predictors that you are using. Are there large amounts of missing data? Is any routine quality check done across databases? Additionally, some details about whether these variables are all clinician-reported, or if some are gathered through patient self-report would be helpful*.

The QICDSS 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 Quebec Public health Institute (INSPQ). Information is only from administrative (i.e. age, hospital or outpatient contact dates) and clinician reported (i.e. diagnoses). Validation of QICDSS physical diagnoses has been achieved by chart reviews and by outcomes for QICDSS psychiatric diagnoses. The QICDSS has been exploited over the past decade by a network of INSPQ officers and academic researchers, many are co-authors, on the characteristics of patients receiving rare psychiatric interventions, but also on personality disorders, schizophrenia and substance use disorders in relation to mortality, including suicide (last paragraph of page 4 and first paragraph of page 6).

3. *If possible, could you add an appendix with your qualitative interview guide to provide the reader with additional context*?

A supplementary file has been included.

4. *A comment or two about anticipated use cases of how policy makers and/or decision makers are hoping to use these models would be helpful – is it to predict yearly suicide rates and*

have initiatives/ campaigns accordingly? Is there the ability to make regional predictions (i.e. more targeted suicide prevention initiatives)?

The application of the prediction models for decision making has been described (third paragraph, page 10).

Responses to rviewer#2's comments

1. *There was some ambiguity in the definition of suicide. The author did not describe the difference between non-Suicidal Self-Injury and suicidal Behaviour, single and Multiple Suicide Attempters, or suicide Attempters and suicide Ideators, whicn will have an enormous influence in the results of models.*

As described in the Methods, the outcome is death by suicide; it is not suicidal behaviors. Therefore, we the description about specific suicidal behaviors is not provided.

2. *"Individual, programmatic, systemic and community factors (see Appendix I) measured five years prior to the suicide events were used to develop the risk predictive algorithms. " The sentence is very unclear. A detailed explanation is needed on the definition of suicide events, and how they can be measured.*

Please see response to reviewer #1's comments#1.

3. *The educational level, income level, employment status and marital status were also associated with the risk of suicide. The risk predictive models for population risk of suicide appear not appropriate without considering these factors.*

We agree that these are important variables. As described in the Methods, we used provincial health administrative database to develop the prediction models. Canada has a public funded healthcare system, and health services are delivered by provinces. Each province has its own health administrative database that contains routinely collected data about health services provided by physicians, procedures, diagnoses, and dates of services. Unlike population health surveys, the health administrative database does not collect information about education, income, employment, and marital status.

4. *The author used a nested case-control study, but the method seems to be inappropriate for this study. In the nested case-control study, cases of a disease that occur in a defined cohort are identified and, for each, a specified number of matched controls is selected from among those in the cohort who have not developed the disease by the time of disease occurrence in the case. In this study, the author used routinely collected health administrative data, without follow-up. Suicide cases and control group were not selected from a same group of samples matching certain conditions. Thus, this study seems to be more like a retrospective study rather than a nested case-control study.*

Thanks for pointing out the error. We have modified the design as case-control study design throughout the text.

Responses to reviewer#3's comments

1. *A first critical point is that the study design is not clearly described. A nested case-control study is a case-control study nested within a cohort (or within a case-cohort). Once the cohort is clearly defined, a number of controls are selected for each individual case, or for the*

number of cases occurring in a certain time-unit (day, week, month, year), using a certain case-control ratio.

Thanks for pointing out the error. We have modified the study design as case-control study throughout the text.

2. *The cohort is not clearly described in the paper, and is not clear which administrative information was used to create the cohort. Information is needed on mortality, birth, immigration and emigration in / out of the population in order to create the retrospective cohort. Which information was used? It would be interesting to know why was the period 2002-2010 was chosen*.

This is a case-control study. We included all death by suicide cases occurred between January 1st 2002 to December 31st 2010, and 1% random sample of living individuals. The predictors are exposures to risk factors (e.g., diagnosis, service use, etc.) in the past 5 years before the index date (see Study Design and Predictors sections on page 5 & 6). We have described how mortality (death by suicide) was ascertained (last paragraph, page 4). We used the data from the period 2002-2010 as the training data; the data collected in 2011 and afterwards will be used for validation.

3. *The authors state that (page 6, line 20) "the control group was a 1% random sample of living individuals in Quebec each year". That does not seem to be conform a nested case-control study. Were controls allowed to be selected more than once across years? Were future cases eligible? Critically, why sample 1% of controls each year? Why not a number proportional to the number of cases in each year?*

Selecting 1% of controls was an arbitrary decision, given the large number of population in the province (over 8 million people). We agree that using the number proportional to the number of cases in each year is another viable option.

4. *A second critical point is the clarity and rationale for the proposed model development. Both logistic regression and machine learning are proposed, and it seems that they serve different objectives. Is that correct? It is not clear (at least, not early on in the paper) which objectives exactly are reached by which techniques, and why different techniques are needed. Machine learning is proposed to be used to enable simulations of policy changes on suicide risk rates, while logistic regression seems to be proposed to be used to forecast future suicide risk based on past data. Why this differentiation in techniques according to different objectives?*

The use of statistical and machine learning techniques is for the same objective – developing sex-specific prediction models to be used by policy and decision makers. Using both techniques, we may compare which approach performs better in predicting population suicide risk and is more feasible to implement. We have clarified this in the revised manuscript (third paragraph, page 6).

5. *A third critical point is that the use of sampling weights is not clear. Why are these sampling weights needed exactly? How are they calculated? Critically, in which analysis (and how) are they applied? Also, it seems to me that class imbalance is an equally important issue but this is nowhere addressed.*

In this case-control study, we included all suicide cases and only 1% of controls. Therefore, the cases and controls don't have an equal probability of being selected. Therefore, in this analytic sample, the proportion of suicide is much higher than that in the target population. As such, for the developed models to be used in the real population (not in the selected sample), weights (the probability of being

selected) should be used in the analysis so that the sample is representative of the target population (third paragraph, page 6). The methods for addressing data imbalance has been added in the paper (see fifth paragraph, page 6)

6. *A fourth critical point refers to what the authors state in the paragraph on Page 8 (paragraph starting at lines 25-26). I do not clearly see the difference between individual-level models and population-level models that the authors introduce here. They state "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 hospitalization 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." I think that this is not unique to population-level models. An individual-level (clinical) model can be developed using data available up to time x (e.g., a certain healthcare visit) and the outcome can be situated at time y, with y being close (days-weeks) to far (months-years) from time x. Any changes in the risk factors of the patient between x and y will not be taken into account by the model.*

We totally agree with this reviewer that this is not unique to population-level models. The risk of developing the outcome at time y is not only dependent on one's exposures at time x and before time x, but conditional on the changes in the exposures between time x and y. However, what is challenging here for an individual-level (clinical) model to follow this principle is the way of model validation. To validate an individual model in a validation data of 1000 or 10,000 people, it would be very challenge (if not possible) to simulate how each individual's exposure will change between time x and y. Whereas for population risk model (or a synthetic model), it is feasible to simulate how demographic composition and health service use in a health region may change between time x and y based on population census and health administrative data, and thereby, to project what may happen in the future in specific health regions or communities. Predicting individual and population risk is very new, and the methodology is not well studied. I certainly welcome and look forward to hearing if this reviewer has other novel approaches for validating prediction models, irrespective of individual and population risk, that may account for the changes between time x and y.

7. *Abstract: the authors introduce the idea of population risk versus individual risk, but mention nowhere that the objective of the study is to "forecast population risk of suicide at the health region level", This should be made clear.*

We have clarified this in the Abstract.

8. *Introduction (page 4, line 6): the authors state that "Suicide has a complex etiology and is a result of the interaction among the risk and protective factors at the individual, healthcare system, and population levels.(3–5)". The cited papers do not represent key papers (e.g., systematic reviews, meta-analyses) on the various types of risk and protective factors for suicide, while many of those key papers are available in the literature.*

We have added more literature including systematic reviews and meta-analysis (second paragraph, pae 3).

9. *Methods - Model development and validation - Data sources (page 5, line 40): while the suicide database and CANUE data are fairly well described, there is no description of the Ministry of Health and Social Services (MSSS) public financial reports. A clear description of all used data should be provided.*

The Ministry of Health and Social Services public Financial reports include the five health administrative databases from which the suicide data were extracted (last paragraph, page 4).

10. *Methods - Model development and validation – Predictors (page 6, line 28): The authors state that "The selection of candidate predictors are determined by content knowledge (i.e., known relationships between suicide or suicide behaviors and individual and local area level quality of health care), feasibility of routine data collection, clinical utility and policy relevance". No literature is cited to back up this statement.*

Our research team is multidisciplinary, including clinical psychiatrists with expertise in suicide treatment and prevention, psychiatric epidemiologists, mental health services research, computer scientists, and mental health policy and decision makers. A large number of variables were extracted from the health administrative data. The initial candidate predictor selection was done through team meetings, based on our content knowledge about the relationships between suicide and suicide behaviors and the variables, the perceived clinical utility, and policy relevance. Although machine learning analysis such as LASSO may also be helpful in selecting candidate predictors, to make the models useful for policy and decision makers, we cannot entirely rely on machine. We decided to take a balanced approach with involvement from both human and machine in the process of predictor selection (second paragraph, page 6).

11. *Methods - Model development and validation - Machine learning (page 7, line 7): The authors state that "From that, we will make simulations of changes coming from policies by modifying the population composition for reflecting the effect of policies change (e.g., mental health diagnoses, socio-economic factors, health system resources allocation) and evaluate their effect on the suicide rates predicted, comparing these with rates obtained with the current population and population modified differently." This was not stated as an objective previously. It should be made more clearly how this will be done, and why machine learning is needed for this.*

What the reviewer cited here is a description of model simulation for estimating intervention effects, which is the application of the developed model. We have removed this sentence to be consistent with the stated objectives.

12. *Model development – logistic regression: Why is a correlation analysis among variables selected by the LASSO regression needed? LASSO will eliminate them. See the paper by Cox et al. (2020) for a discussion of this topic.*

Theoretically, LASSO may penalize the coefficients of unpredictive variables to zero. It is a very useful tool. However, there are still variables that are strongly correlated after LASSO. Therefore, a correlation analysis will be carried out to examine the possibility (third paragraph, page 7).

13. *Methods - Model development and validation (page 7, line 41): The authors state that "That accounted for clusterings at the health regions". How exactly will this be done?*

We will accomplish this by including a health region variable (third paragraph, page 7).

14. *Methods - Model development and validation (page 7, line 41): the authors state that "Backward selection method will be used to identify the model with the best calibration and discrimination". Backward selection, as I know it, is a method to select variables to include in a model. The use of a "backward selection method" needs to be clarified in the context of model selection. Especially since LASSO was earlier described as a predictor selection method.*

After LASSO, there may be a large number of candidate predictors. Backward selection method will be used to eliminate unpredictive variables and to identify the model with the best calibration and discrimination. We have clarified this in the revised paper (third paragraph, page 7).

15. *Methods - Model development and validation (page 8, line 11): The authors state that "First, for each predictor, the proportions of individuals within each category of that predictor in the initial modeling will be computed, separately by regions. For instance, if hospitalization 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." This seems like a rather elaborate way to explain that aggregate values by health region of predictor variables will be calculated. Is this correct?*

The reviewer is correct. To make prediction at the health region level, one needs to estimate the aggregate values of the predictors for each health region so that the synthetic models can make the prediction based on the aggregate values.

16. *Abstract: the author mention "synthetic estimation models". It is not clear what is meant by "synthetic".*

The limited space in Abstract will not allow detailed description about synthetic models. Therefore, we have removed "synthetic estimation" from the Abstract.

17. *Methods - Model development and validation (page 4, line 52): the authors state "and the indicators at each level may be classified into the broad categories of input and process". It is not clear what is meant by "input and process".*

We have removed this sentence from the paper.

We hope that the revision is satisfactory. If you have more questions, please contact me.

VERSION 2 – REVIEW

REVIEWER	Sequeira, Lydia University of Toronto Institute of Health Policy Management and Evaluation
REVIEW RETURNED	05-Dec-2022

GENERAL COMMENTS	Thank you for addressing the comments and updating your methods & limitations section appropriately.

REVIEWER	Yin, Honglei Southern Medical University
REVIEW RETURNED	22-Nov-2022

GENERAL COMMENTS	Accept

REVIEWER	Mortier, Philippe Hospital del Mar Institute for Medical Research, Health Services Research Group
REVIEW RETURNED	04-Dec-2022

GENERAL COMMENTS	After going through the revised version of this protocol paper, these main issues remain unaddressed in my opinion: *** The study design. The authors changed the study to a case-control study, in stead of a nested case-control study. However, it is still not clear why 1% of living individuals are selected in each year as controls. In addition, it is not clear where exactly (from which lists) controls are sampled from. Related to this, it is not clear if controls are allowed to be selected more than once across years, and if future cases are eligible as controls. It is therefore still not clear whether the study design is a nested case-control study, a case-cohort study, or a case-control study. For more information that may be helpful in clarifying the study design, see Chapter 8 of Lash, Vanderweele, Haneuse, and Rothman's textbook Modern Epidemiology. *** The necessity to use sample weights is still unclear. First, when using a case-control design, estimates of association (e.g., Odds Ratio) do not change in comparison with an estimate of the same correlation using the full cohort or population data. Estimation is just more efficient. Weights are not needed to calculate estimates of association. Second, when using machine-learning based classification algorithms, class imbalance is addressed by methods that try to obtain comparable proportions of cases and controls. So it is not clear why weights are needed when conducting the analyses. The authors should make clear when and why exactly sample weights are needed. *** The strategy the authors propose to adjust for clustering of data in healthcare regions (and associated bias in variance and standard error estimation) is inadequate. Simply including healthcare region as a covariate, as the authors here state, is inadequate. Robust estimators, or multilevel modeling are valid options to adjust variance estimation for clustering. However, since prediction seems to be the main objective of this study, it is unclear why variance estimation (and adjustment for clustering) is needed in the first place.

VERSION 2 – AUTHOR RESPONSE

Responses to reviewer #3's comments:

1. *The study design. The authors changed the study to a case-control study, in stead of a nested case-control study. However, it is still not clear why 1% of living individuals are selected in each year as controls. In addition, it is not clear where exactly (from which lists) controls are sampled from. Related to this, it is not clear if controls are allowed to be selected more than once across years, and if future cases are eligible as controls. It is therefore still not clear whether the study design is a nested case-control study, a case-cohort study, or a case-control study.*

The controls were the living individuals between January 1st, 2002 and December 31st, 2010, who were selected from the Quebec physician claim database, which is one of five databases the INSPQ assembled for the suicide database (see the description about the data sources). Essentially, this database convers a majority of people living in the province. 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. We have specified these in the revised manuscript (see the paragraph for study design, page 5).

2. *The necessity to use sample weights is still unclear. First, when using a case-control design, estimates of association (e.g., Odds Ratio) do not change in comparison with an estimate of the same correlation using the full cohort or population data. Estimation is just more efficient. Weights are not needed to calculate estimates of association. Second, when using machine-learning based classification algorithms, class imbalance is addressed by methods that try to obtain comparable proportions of cases and controls. So it is not clear why weights are needed when conducting the analyses. The authors should make clear when and why exactly sample weights are needed*.

It is well acknowledged in epidemiology that case-control study design produces a biased sample. The bias is due to the fact that the proportion of cases in the sample is not the same as the population of interest. One method for addressing this limitation when developing predictive models using case-control data is weighting. Therefore, sample weights were used in logistic regression modeling (see page 6). This approach has been used in previous studies.

3. *The strategy the authors propose to adjust for clustering of data in healthcare regions (and associated bias in variance and standard error estimation) is inadequate. Simply including healthcare region as a covariate, as the authors here state, is inadequate. Robust estimators, or multilevel modeling are valid options to adjust variance estimation for clustering. However, since prediction seems to be the main objective of this study, it is unclear why variance estimation (and adjustment for clustering) is needed in the first place.*

We agree with this comment, and have removed the sentence about adjusting for clustering effect by health regions from the manuscript (page 7).