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)	Safe opioid prescribing: a prognostic machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada
AUTHORS	Sharma, Vishal; Kulkarni, Vinaykumar; Eurich, Dean; Kumar, Luke; Samanani, Salim

VERSION 1 – REVIEW

REVIEWER	Zhou, Lili AbbVie Inc, Health economics and outcome research
REVIEW RETURNED	10-Oct-2020

GENERAL COMMENTS	This manuscript entitled “Safe opioid prescribing: a machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada” aimed to predict a 30-day risk of hospitalization, emergency department visit, or mortality following an opioid prescription dispensation among the opioid users residing in Alberta, Canada, using machine learning techniques. This study used different types of machine learning models to maximize the prediction power. Overall, this is a relevant topic in the field given the high opioid prescribing in Canada. While the reviewer appreciates the importance of this work, there are many notable major concerns in the study design, methods and other areas in the manuscript needed to be clarified in order to ensure the validity of the findings. Areas could be improved to strengthen the manuscript and study are listed below. Abstract 1. (Page 4, line 35-36): Authors described “a post-test probability of 13%, up from the pre-test probability of 1.6%” in the results of the abstract, but the reviewer could not locate where 13% was mentioned in the main text or in the table in this manuscript. Could authors clarify that? Introduction 2. (Page 6, line 5): “Canada has among the highest rates of opioid prescribing in the world”. This sentence has a grammar error. Suggest changing to “Canada is among the highest …..” 3. (Page 6, line 28-30): “both population-level monitoring ……..most often based on prescribing guidelines”. The font size of this sentence is smaller than the rest of texts. Please correct it. 4. (Page 7, line 16-26): Authors described a previous study (citation number 10) that applied ML techniques to predict overdose risk in opioid patients, but reviewer is not clear about what incremental value that this study would provide compared with the previous study. Suggest authors adding more discussions to clarify the research gaps, as well as the necessity of this study to emphasize the importance of this study.

	5. (Page 7, line 28-30): The study objective was to predict the 30-day risk of hospitalization, emergency visit, or mortality following an opioid prescription among patients residing in Alberta, Canada. Could authors clarify why predicting a 30-day risk rather than a shorter or long period? 6. (Page 8, line 42-46): The authors briefly described that the data "Physician Visits/Claims (Alberta Health)" included ICD codes associated with the claims, procedures and billing information. Could authors clarify whether the physician visits information includes all type of health settings such as inpatient, outpatient, and skilled nursing facility, or just inpatient and emergency department visits? 7. (Page 9, line 22-25): The included comorbidity history using ICD-based Elixhauser score categories. Based on the eTables 3-4, the Elixhauser comorbidity index only included 28 conditions. Since the primary purpose of this study was prediction, why didn't authors identify all the comorbidities and use them in the model as it may further improve the prediction performance? Elixhauser comorbidity index was developed by the Healthcare Cost and Utilization Project. All the ICD diagnosis codes used in the original Elixhauser index were ICD-9-CM and ICD-10-CM codes, different from the WHO ICD-9/ICD-10 codes used in this study. Did authors adapt the ICD-CM codes to WHO ICD codes? If so, please list the adapted codes in a table in appendix. 8. (Page 9, line 30-32): Authors mentioned that the medication use features include "concurrent use with benzodiazepines, number of opioid and benzodiazepine dispensations", but based on eFigure 1, the drug utilization was captured within 30-day prior to opioid dispensation. Could authors clarify how concurrent use with benzodiazepines is defined? 9. (Page 9, line 32-37): Authors stated that "we used data from 30 days to 5 years before the opioid dispensation to generate model features (eFigure 1 in Supplement)". Could authors further clarify why using different time windows to capture features of drug utilization, healthcare utilization, and comorbidity? In addition, don't authors agree that looking back 5-year to identify comorbidities too long? 10. (Page 10, the first paragraph): Authors described many analyses that they would do to test the model performance of discrimination and calibration. The reviewer suggests breaking this paragraph into two, with one describing discrimination and the other describing calibration. The current paragraph is lengthy and easy to lose readers. Also, suggest adding some contents to explain what discrimination and calibration represents? What is the difference between then? and why both are needed to evaluate the model performance? Results 11. (Page 11, line 37-51): Authors described the number of patients eligible for the analysis as well as the number of patients with the outcomes in 2017 and 2018, respectively, between line 36-39. Later, authors further also described the number of opioid prescriptions as well as the number of opioid prescriptions associated with the outcomes. Could authors clarify the observation unit in this study? A patient-level or prescription-level? In training and validating the model, can one patient be counted only once or would be counted multiple times if he/she has 2 or more opioid prescriptions? Please clarify the analysis unit. 12. (Page 11, line 46-51): The results showed that "in 2017, 2.03% (n=45,757) of opioid prescriptions were associated with the outcomes; in 2018, the estimate was 1.6% (n=31,392)". Based on

	the results, the outcomes were relatively rare. Did the authors use any approach to balance the number of two classes (outcome vs non-outcome) in the model? An imbalanced classification would cause a problem in training and validating the model. Please read the article below. https://machinelearningmastery.com/what-is-imbalanced-classification/ 13. (Page 13, last paragraph): Authors described the discrimination performance using the US and Canadian guideline recommendation and presented the results in Table 2. However, it looks like some of performance measures such as high opioid dose (>120 PME/day for 90+ days), multiple doctors (>4), and mental disorder OR substance abuse OR OME/day >90 are measured in a longer period other than 30-day. How did authors make a prediction in using these measures? Suggest authors adding more descriptions regarding how the prediction was made in using the clinical guidelines. Discussion 14. (Page 15, second paragraph): Authors compared this study with previous studies that applied ML techniques to predict opioid-related adverse outcomes. Again, based on the discussions, it is not clear what incremental value does this study provide. Suggest adding more discussions to clarify the importance of this study and to emphasize the necessity of this study. 15. (Page 16, last paragraph): Authors concluded that this study could support PMPs as the ML models could better identify the high risky patients. However, authors did not discuss about the clinical implication of the ML models. How PMPs could use the models to predict the risk among the opioid users? Suggest authors adding more discussions about the clinical implications of this study to further emphasize the importance of this study. Tables and figures 16. It is hard to read the Figure 1 and eFigure 3. Please increase the resolution. 17. Only 11 features were included in eFigure 2. Could authors clarify why other features are not presented? In addition, did authors use a regular logistic regression without regularization? If yes, why don't authors use a regularized logistic regression (e.g., Lasso) as it may further improve the model performance? Please clarify.

REVIEWER	Deng, Hao Massachusetts General Hospital, Department of Anesthesia Critical Care and Pain Medicine
REVIEW RETURNED	17-Dec-2020

GENERAL COMMENTS	The reviewer appreciated that the authors' efforts to comply with the TRIPOD guideline in this submitted manuscript. The topic is certainly interesting and the methods used are solid. There some issues need to be addressed and are presented below: 1. P5L22 Abstract: The authors stated that the 2018 data is "independent validation sets", which is inaccurate. The 2017 validation set can be considered as independent validation set if random split process was performed as designed. But all 2018 data are temporally correlated with 2017 data, making it depended data. The reviewer suggests the authors to make a minor change to remove the word "independent" in describing 2018 validation set.

	2. P8L28-40 Introduction: The authors stated that the objective is develop and validate ML algorithms for predictions, and hypothesized that the newly developed models would outperform the current guidelines. However, no formal statistical tests were planned or performed. The reviewer recommends that the authors can consider either revise the language and tests to compare models to a fixed criteria (e.g., 80%) or perform formal comparisons with current guideline performance. For instance, comparing two AUC-ROC curves. 3. P8L55 Exclusion criteria: The authors set the exclusion criteria to exclude cancer patients, palliative care patients, and pregnant women from the analysis. The exclusion criteria are reasonable. The reviewer is curious about whether the current guideline predictions had excluded these patients in modeling as well. Please clarify and provide related information. 4. P9L54 methods: Please provide the names of key packages for computing and modeling such as PyTorch, TensorFlow or XGBoost. 5. P11 statistical analysis: Please include a power statement in method sections. 6. P11 statistical analysis: please include a missing data statement in method sections. This is important and the reviewer is interested to know what the missing patterns were and how the authors addressed missing data in modeling processes. 7. P13L27 results: Please reconsider the word use of “predictive power”, because power is a special term in statistics as stated in previous sections. 8. P13L45 results: The reviewer suggests the authors to re-scan the manuscript to carefully choose their words when describing predictive performance. For instance, quote “...the highest impact for...” is not accurate. When we talk about impact, we are talking about inferences, while the authors were really talking about “variable importance” from the model outputs. Also, if the authors intend to interpret the prediction models in a hypothesis testing framework, it is recommended to avoid word usage such as “tendency”. 9. Not much detail was provided regarding NNs. Please provide necessary information regarding NN model construction. Is that a RNN, simple feed-forward or something else? Please clarify. The same for Bayesian modeling. How did you define the prior? Where are the uncertainty measures? 10. While the reviewer agrees that it is interesting to view horizontal comparisons among different categories of methods for modeling. It appears to me that the authors were only fitting these models in out-of-box settings. Just a personal opinion, the authors can consider to customize a little more on the modeling process such as using validated regularization methods such as Elastic-net/Lasso to boost the performance of regression models or even SVM. But for this particular manuscript, I only suggest the authors to consider mention this point as a potential limitation.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Dr. Lili Zhou, AbbVie Inc

Comments to the Author:
This manuscript entitled “Safe opioid prescribing: a machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada” aimed to predict a 30-day risk of hospitalization, emergency department visit, or mortality following an opioid prescription dispensation among the opioid users residing in Alberta, Canada, using machine learning techniques. This study used different types of machine learning models to maximize the prediction power. Overall, this is a relevant topiin the field given the high opioid prescribing in Canada. While the reviewer appreciates the importance of this work, there are many notable major concerns in the study design, methods and other areas in the manuscript needed to be clarified in order to ensure the validity of the findings. Areas could be improved to strengthen the manuscript and study are listed below.

Abstract
1. (Page 4, line 35-36): Authors described “a post-test probability of 13%, up from the pre-test probability of 1.6%” in the results of the abstract, but the reviewer could not locate where 13% was mentioned in the main text or in the table in this manuscript. Could authors clarify that?
Response:
Table 1 has the statistic in question. We rounded to 13% (from 13.38, 13.45) but will clarify the Abstract as recommended by the Reviewer:
“The top 5-percentile of predicted risk for the XGBoost and logistic regression classifiers captured 42% of all events and translated into post-test probabilities of 13.38 and 13.45%, respectively, up from the pre-test probability of 1.6%.”

Introduction

2. (Page 6, line 5): “Canada has among the highest rates of opioid prescribing in the world”. This sentence has a grammar error. Suggest changing to “Canada is among the highest …..”
Response:
We will change the sentence according to Reviewer recommendation:
“Canada is among the countries with the highest rates of opioid prescribing in the world,”

3. (Page 6, line 28-30): “both population-level monitoring ……..most often based on prescribing guidelines”. The font size of this sentence is smaller than the rest of texts. Please correct it.

Response:
Font size corrected

4. (Page 7, line 16-26): Authors described a previous study (citation number 10) that applied ML techniques to predict overdose risk in opioid patients, but reviewer is not clear about what incremental value that this study would provide compared with the previous study. Suggest authors adding more discussions to clarify the research gaps, as well as the necessity of this study to emphasize the importance of this study.

Response:

We amended the Introduction:

"In their validation sample, they found that the DNN (deep neural network) and GBM (gradient boosting machine) algorithms carried the best discrimination performance based on estimated c-statistics and that the ML approach out-performed the guideline approach in terms of risk prediction; neural networks have little interpretability and are not necessarily better at predicting outcomes when trained on structured data 18. This study relied on c-statistics to evaluate their ML models and did not emphasize other performance metrics required to assess clinical utility that are recommended by medical reporting guidelines 11,13,19,20. It also did not address the important issue of ML model interpretability 21. Reporting informative prognostic metrics is needed to better understand the capabilities of ML classifiers if health departments and PMPs are to incorporate them into their decision-making processes.

The objective of our study was to further develop and validate ML algorithms (beyond just DNN) to predict the 30-day risk of hospitalization, emergency visit and mortality for a patient in Alberta, Canada at the time of an opioid dispensation using administrative data routinely available to health departments and PMPs and evaluate them using the above referenced reporting guidelines. We also analyzed feature importance to provide meaningful interpretations of the ML models. Comparing discrimination performance (area under the receiver operating characteristics curves), we hypothesized that the ML process would perform better than the current guideline approach for predicting risk of adverse outcomes related to opioid prescribing. "

5. (Page 7, line 28-30): The study objective was to predict the 30-day risk of hospitalization, emergency visit, or mortality following an opioid prescription among patients residing in Alberta, Canada. Could authors clarify why predicting a 30-day risk rather than a shorter or long period?

Response:
We used 30-day risk because potential end users expressed interest in 30-day risk and that 30 days is a routinely used endpoint in health services research in Canada, resource allocation and in risk prediction models in general. We will cite a reference in the manuscript regarding this issue.

6. (Page 8, line 42-46): The authors briefly described that the data "Physician Visits/Claims (Alberta Health)" included ICD codes associated with the claims, procedures and billing information. Could authors clarify whether the physician visits information includes all type of health settings such as inpatient, outpatient, and skilled nursing facility, or just inpatient and emergency department visits?

Response:
Claims include claims submitted from all settings (e.g., outpatient office visits, ED, inpatient, etc).. We will edit the manuscript as recommended by the Reviewer as follows:
"4) Physician Visits/Claims (Alberta Health): all claims from all settings (e.g., outpatient, office visits, emergency departments, inpatient) with associated date of service, ICD code , procedure and billing information."

7. (Page 9, line 22-25): The included comorbidity history using ICD-based Elixhauser score categories. Based on the eTables 3-4, the Elixhauser comorbidity index only included 28 conditions. Since the primary purpose of this study was prediction, why didn't authors identify all the comorbidities and use them in the model as it may further improve the prediction performance?
Elixhauser comorbidity index was developed by the Healthcare Cost and Utilization Project - All the ICD diagnosis codes used in the original Elixhauser index were ICD-9-CM and ICD-10-CM codes,

different from the WHO ICD-9/ICD-10 codes used in this study. Did authors adapt the ICD-CM codes to WHO ICD codes? If so, please list the adapted codes in a table in appendix.

Response:
The Reviewer is correct. There are only 28 co-morbidities in eTables 3,4 because we collapsed uncomplicated, complicated hypertension and diabetes into 1 each, thus reducing the Elixhauser categories to 28 from 30. We also added injury and poisoning to the list as these are comorbidities of concern in an opioid population. We will add a footnote to eTables 3 and 4 to reflect the Reviewer's comments as follows:
"** Complicated, uncomplicated diabetes and hypertension were collapsed into one category each"

To code the Elixhauser co-morbidities, we used what others have, namely, the Hude Quan paper, which used ICD-10 and ICD-9 CM and who has validated these codes for the Canadian context. We will add this reference to the Methods section to reflect this to clarify as recommended by the Reviewer.

8. (Page 9, line 30-32): Authors mentioned that the medication use features include "concurrent use with benzodiazepines, number of opioid and benzodiazepine dispensations", but based on eFigure 1, the drug utilization was captured within 30-day prior to opioid dispensation. Could authors clarify how concurrent use with benzodiazepines is defined?

Response:
Concurrent use is defined as 7 days of cumulative concurrent use in the previous 30 days prior to opioid dispensation based on days' supply
We added this definition to eTable 5 to clarify the Reviewer's point.

9. (Page 9, line 32-37): Authors stated that "we used data from 30 days to 5 years before the opioid dispensation to generate model features (eFigure 1 in Supplement)". Could authors further clarify why using different time windows to capture features of drug utilization, healthcare utilization, and comorbidity? In addition, don't authors agree that looking back 5-year to identify comorbidities too long?

Response:
-We used the timelines 30 days to 5 years based on availability of data to maximize our predictive capabilities in the models. As is typical in health services research, we used all available data to determine and characterize comorbidities in our population. Co-morbidity history is routinely longer term, permanent ailments a patient may have (e.g., congestive heart failure). Chronic co-morbidities we used are lifelong (e.g., history of CVD) and using shorter time frames would miss important co-morbidities that could be present in a patient, influence prescribing, and potential have predictive power in our models. For example, see this paper in BMJ
-Conversely, we also used a number of shorter windows (e.g., 180 to 30 days) to reflect immediate nature of risk of outcomes. For example, use of a prescription drug 6 months ago is unlikely to influence a hospitalization today; however, a prescription dispensed within the last 30 days certainly could influence the risk of a hospitalization.
- As the author is likely aware, there is no consensus in the literature as to what time windows to use. The paper we cited used 3 month rolling windows which was an arbitrary decision. This would be consistent with our shorter time windows but we also believe this approach would miss important comorbidities that patients may have which occurred in the past.
-Last, and maybe most importantly, one of our aims was to make the models clinically relevant and interpretable. Most clinicians would approach a patient in a similar manner with respect to

care. They would look at the entire history of comorbidities of a patient and then the more immediate factors in deciding on the need for a therapeutic as well as risk in patients.

To reflect the Reviewer's comments, we amended the Methods section as follows:
"Depending on the potential predictor and data availability, we used data from 30 days to 5 years before the opioid dispensation to generate model features (eFigure 1 in Supplement); 30 days was used to reflect the immediate nature of the risk and 5 years to fully capture co-morbidities. This approach aligns with how health providers would assess patients using the entire history of co-morbidities and then the more immediate factors in deciding on the need for a therapeutic as well as risk in patients. "

10. (Page 10, the first paragraph): Authors described many analyses that they would do to test the model performance of discrimination and calibration. The reviewer suggests breaking this paragraph into two, with one describing discrimination and the other describing calibration. The current paragraph is lengthy and easy to lose readers. Also, suggest adding some contents to explain what discrimination and calibration represents? What is the difference between then? and why both are needed to evaluate the model performance?

Response:
Reviewer is correct, this section will be clarified as follows:
" First, we trained commonly used[13,29] ML algorithms (eAppendix in Supplement) and tuned model hyperparameters using k-fold (k=5) cross validation to address model overfitting[13,30]. As is common in ML validation studies[10,13], we reported model discrimination performance (i.e. how well a model differentiates those at higher risk from those at lower risk)[11] using area under the receiver operating characteristic curve (AUROC; c-statistic). We then stratified the two ML models with the highest c-statistics into percentile categories according to absolute risk of our outcome, as was done in previous studies[10,31]. We also plotted AUROC[11] and precision-recall curves (PRCs)[32].
Because discrimination alone is insufficient to assess ML model prediction capability, we assessed a second necessary property, namely, calibration (i.e. how similar the predicted absolute risk is to the observed risk across different risk strata)[11,35]. Using the two ML models with the highest discrimination performance discussed above, we assessed calibration performance on the 2018 data by plotting observed (fraction of positives) vs predicted risk (mean predicted value). Using these two ML classifiers , we analyzed the top 0.1, 1, 5, and 10 percentiles of predicted risk by the number of true and false positives, positive likelihood ratios (PLR)[20], post-test probabilities, and number needed to screen. We also performed a simulation of daily data uploads for 2018 Quarter 1 to view the predictive capabilities if a ML risk predictor were to be deployed into a monitoring workflow.
For the XGBoost and logistic regression classifiers, we reported feature importance[29] and plotted PRCs that compared all dispenses to those within the top 10 percentiles of estimated risk. As well, for the XGBoost classifier, we described feature impact on model outcome using SHAP values[33,34] to add an additional layer of interpretability.
Finally, we compared ML risk prediction (the two ML models with highest discrimination performance) to current guideline approaches as others have[10], using the 2019 Centers for Medicare & Medicaid Services opioid safety measures[36] and the 2017 Canadian Opioid Prescribing Guideline[3]. We also compared the discrimination performance of different logistic regression classifier models using various combinations of features derived from their respective databases: 1) demographic and drug/health utilization features from PIN and 2) co-morbidity features derived from DAD, NACRS and Claims.
"

Results

11. (Page 11, line 37-51): Authors described the number of patients eligible for the analysis as well as the number of patients with the outcomes in 2017 and 2018, respectively, between line 36-39. Later, authors further also described the number of opioid prescriptions as well as the number of opioid prescriptions associated with the outcomes. Could authors clarify the observation unit in this study? A patient-level or prescription-level? In training and validating the model, can one patient be counted only once or would be counted multiple times if he/she has 2 or more opioid prescriptions? Please clarify the analysis unit.

Response:
We will clarify our manuscript with the Reviewer's recommendations:
-Observation unit (ML models were trained with) is opioid dispensations.
-yes, a patient can be represented in more than 1 instance within training and validation sets, but they cannot be in both sets
We will amend the Methods section "Measures and Outcomes" as follows:
"Measures and Outcome
ML models were trained on a labelled dataset in which the observation/analysis unit was an opioid dispensation."

We amended the Methods-Statistical Analyses and ML Prediction Evaluation section as follows:
"We randomly divided the patients in the 2017 portion of our study cohort into training (70%) and validation (30%) sets[13] by patients and dispensations such that no patients in the training set were in the validation set. "

12. (Page 11, line 46-51): The results showed that "in 2017, 2.03% (n=45,757) of opioid prescriptions were associated with the outcomes; in 2018, the estimate was 1.6% (n=31,392)". Based on the results, the outcomes were relatively rare. Did the authors use any approach to balance the number of two classes (outcome vs non-outcome) in the model? An imbalanced classification would cause a problem in training and validating the model. Please read the article below.
https://machinelearningmastery.com/what-is-imbalanced-classification/

Response:
Reviewer is correct in this statement.
- We tried oversampling (randomly repeating minority class) and under sampling (sub sampling within majority class) approaches, but they changed the prevalence of the outcome in the population which is not ideal for a model used in real world. Hence we relied on class weightage - for the supported algorithms we used balancing class weights

We will clarify this point and cite the website provided as follows in the Methods section-Measures and Outcomes:
"We anticipated that our defined outcome would be a rare event, leading to a class imbalanced dataset[31]. To address this, we relied on specifying balanced class weightage for supporting algorithms; other approaches were not deemed suitable (e.g., randomly repeating minority class) and under sampling (sub-sampling within the majority class) resulted in changes in outcome prevalence."

13. (Page 13, last paragraph): Authors described the discrimination performance using the US and Canadian guideline recommendation and presented the results in Table 2. However, it looks like some of performance measures such as high opioid dose (>120 PME/day for 90+ days), multiple doctors (>4), and mental disorder OR substance abuse OR OME/day >90 are measured in a longer

period other than 30-day. How did authors make a prediction in using these measures? Suggest authors adding more descriptions regarding how the prediction was made in using the clinical guidelines.

Response:
-For the CMS and CAD guidelines, we re-engineered the data to assess the guideline based approach and used timelines specified within the guidelines. The US CMS guidelines identify a timeline of 90 or more days at >120 OME and concurrent use for 30 days or more. CAD opioid guidelines do not specify timelines.
We will add footnotes to Table 2 and reference/cite CMS and CAD guidelines in the Methods section to provide more details on how we assessed guidelines as follows as recommended by the Reviewer:
" *The Canadian guidelines do not specify timelines. >90 OME was determined by taking the average daily OME over the 30 days prior to dispensation
**The CMS guidelines specify a timeline of 90 or more days at >120 OME and concurrent use of opioids and benzodiazepines for 30 days or more"

Discussion

14. (Page 15, second paragraph): Authors compared this study with previous studies that applied ML techniques to predict opioid-related adverse outcomes. Again, based on the discussions, it is not clear what incremental value does this study provide. Suggest adding more discussions to clarify the importance of this study and to emphasize the necessity of this study.

Response:
Reviewer is correct. We will clarify by adding the piece where our study used informative metrics to assess the clinical utility of our ML classifiers, metrics that others do not emphasize.
We will clarify as recommended in the Discussion:
"We found only one study that used ML approaches to quantify the absolute risk of an event pursuant to an opioid dispensation[10]. Their methodology used rolling 3-month windows for estimating risk and ML model training while we used historic records to estimate 30-day risk. Differences in study population and feature selection may explain why their highest performing ML model was deep learning (neural network classifier) and ours was not. Nevertheless, we were able to replicate their predictive performance using our ML approach as we both showed that ML approaches have higher predictive capabilities than guideline approaches. Both of our studies used predicted percentile risk estimates to identify high risk dispensations and were able to do so with strong discrimination and calibration performance. Furthermore, we emphasized prognostic metrics which are more informative to assess the clinical utility of ML classifiers using pre- and post-test probabilities, something not done in other studies and recommended in medical guidelines[20]. This major aspect of our study, not done previously, is important because any ML classifier that does not increase prognostic information compared to baseline cannot be incorporated into decision making for the purpose of intervening on higher risk instead of lower risk patients. Indeed, another study we found describes how identifying cases in higher predicted risk percentiles using ML methods can be deployed in hospital settings for the purpose of targeted interventions[32] upon discharge, however the effect on outcomes is still to be determined."

15. (Page 16, last paragraph): Authors concluded that this study could support PMPs as the ML models could better identify the high risky patients. However, authors did not discuss about the clinical implication of the ML models. How PMPs could use the models to predict the risk among the

opioid users? Suggest authors adding more discussions about the clinical implications of this study to further emphasize the importance of this study.

Response:
Reviewer is correct. We amended the last paragraph as follows:
"This study suggests that ML risk prediction can support PMPs, especially if readily available administrative health data is used. PMPs currently use population-based guidelines which we, and others, have shown cannot predict absolute individual risk. The ML process allows for model training, validation and deployment to specific settings in which, for the case of PMPs, high risk patients can be identified and targeted for intervention either at the patient or provider level. Moreover, ML classifiers can be retrained over time as changes in populations and trends in prescribing occur and are therefore specific to the population unlike broadly based guidelines. Further research can assess whether implementation of a ML-based monitoring system by PMPs leads to improved clinical outcomes."

Tables and figures

16. It is hard to read the Figure 1 and eFigure 3. Please increase the resolution.
Response:
We increased the size of these figures as recommended by the Reviewer.

17. Only 11 features were included in eFigure 2. Could authors clarify why other features are not presented? In addition, did authors use a regular logistic regression without regularization? If yes, why don't authors use a regularized logistic regression (e.g., Lasso) as it may further improve the model performance? Please clarify.

Response:
-for eFigure2, we only chose to illustrate the most important features out of the many (>200 features) as determined by H20 driverless. As for logistic regression – we used L1-Lasso regression
We amended the eAppendix section on Logistic Regression as follows to clarify:
"We used the sci-kit learn library in our experiments[6] and found L1 regularization to be more effective."
We also added a footnote in the Tables to specify we used lasso regularization:
"Note: Logistic regression used L1 (lasso) parameter regularization"

Reviewer: 2
Dr. Hao Deng, Massachusetts General Hospital

Comments to the Author:

The reviewer appreciated that the authors' efforts to comply with the TRIPOD guideline in this submitted manuscript. The topic is certainly interesting and the methods used are solid. There some issues need to be addressed and are presented below:

1. P5L22 Abstract: The authors stated that the 2018 data is “independent validation sets”, which is inaccurate. The 2017 validation set can be considered as independent validation set if random split process was performed as designed. But all 2018 data are temporally correlated with 2017 data, making it depended data. The reviewer suggests the authors to make a minor change to remove the word “independent” in describing 2018 validation set.

Response:
-Reviewer is correct and we removed “independent”

2. P8L28-40 Introduction: The authors stated that the objective is develop and validate ML algorithms for predictions, and hypothesized that the newly developed models would outperform the current guidelines. However, no formal statistical tests were planned or performed. The reviewer recommends that the authors can consider either revise the language and tests to compare models to a fixed criteria (e.g., 80%) or perform formal comparisons with current guideline performance. For instance, comparing two AUC-ROC curves.

Response:
We amended the Introduction as recommended by Reviewer to read:
“Comparing discrimination performance (area under the receiver operating characteristics curves), we hypothesized that the ML process would perform better than the current guideline approach for predicting risk of adverse outcomes related to opioid prescribing.”

3. P8L55 Exclusion criteria: The authors set the exclusion criteria to exclude cancer patients, palliative care patients, and pregnant women from the analysis. The exclusion criteria are reasonable. The reviewer is curious about whether the current guideline predictions had excluded these patients in modeling as well. Please clarify and provide related information.

Response:
Reviewer makes a valid point in that the guidelines are in the context of chronic non-cancer pain. Pregnant women are not specified in the guidelines
-In our analysis of CAD and CMS guidelines, we excluded pregnant patients and cancer and palliative for both ML and guideline predictive modeling.
We will amend the Methods section to read:
“. Patients were excluded from all analyses if they had any previous diagnosis of cancer, received palliative interventions or were pregnant during the study period (eTable 1 in Supplement) as use of opioids in these contexts is clinically different.”

4. P9L54 methods: Please provide the names of key packages for computing and modeling such as PyTorch, TensorFlow or XGBoost.

Response:
Python (v. 3.6.8), SciKit Learn (v. 0.23.2), SHAP (v.) , XGBoost (version 0.90) –

VERSION 2 – REVIEW

REVIEWER	Zhou, Lili

	AbbVie Inc, Health economics and outcome research
REVIEW RETURNED	07-Feb-2021

GENERAL COMMENTS	This manuscript entitled “Safe opioid prescribing: a machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada” aimed to predict a 30-day risk of hospitalization, emergency department visit, or mortality following an opioid prescription dispensation among the opioid users residing in Alberta, Canada, using machine learning techniques. This study used different types of machine learning models to maximize the prediction power. Overall, this is a relevant topic in the field given the high opioid prescribing in Canada. The authors have well addressed my comments on the study design and methods, but there are many notable minor concerns in the manuscript needed to be clarified in order to ensure the validity of the findings. Areas could be improved to strengthen the manuscript and study are listed below. 1. (Page 5, line 10): The authors claimed that one strength of this study was “The study population is the entire provincial population and is generalizable to other populations in Canada and beyond”. Suggest removing this strength as it is not appropriate from reviewer’s perspective. This study includes the entire population living in Alberta, Canada, so it is confident to say that results are well representative for this province. However, it is not appropriate to say that the results can be generalizable to other population in Canada because this study did not include participants from other regions of Canada and there are lack of evidence indicating that the opioid use pattern and subsequent risks were the same in other regions in Canada as in Alberta. Please also remove the relevant sentences in the discussion section. 2. Could authors list all the ingredients (e.g., hydrocodone, oxycodone) included in opioids in this study in a supplemental table? Did authors also include opioid antitussives used for treatment of cough in the study? Did authors include both oral and parenteral opioids? The previous paper (citation number 10: Lo-Ciganic et al published in JAMA Open) that cited in this manuscript excluded individuals with only parenteral opioid prescription and/or cough or cold medications prescriptions containing opioids as these medications are either for acute pain relief or have low addictive potency. Suggest authors applying the same exclusion criteria in this study. 3. (Page 9, line 37): Authors described the primary outcome of this study a composite of a DRUG-RELATED hospitalization, ED visit or mortality within 30 days of an opioid dispensation. How did authors determine whether the hospitalization, ED visit or mortality was caused by opioid use? Please clarify it. 4. (Page 9, line 47): Authors stated “To address this, we relied on specifying balanced class weightage for supporting algorithms”. Could authors add more discussions of advantages of using this approach over the others that authors did not deem suitable. Also, please cite a reference here. 5. How many exact features were included in the model? I could not find the specific number in the text. Please clarify it. In addition, in my previous comment, I asked why only Elixhauser

	comorbidities were included in the model? Shouldn't including more comorbidity features would further improve the model prediction power? In addition, for the feature category of healthcare utilization, only two features, including number of hospitalizations and number of unique providers, were included. Why not including more health utilization features to increase the prediction power (e.g., number of ED visits)? I understand that authors want to minimize the feature numbers by only including the potential strong predictors, but could authors clarify why only Elixhauser comorbidities and two healthcare utilization features were used? My concern is that authors may miss some important features. Thank you! 6. Authors included a feature of "concurrent use with benzodiazepines". Reviewer asked for the clarification of definition of this feature. In the response letter, authors said that "concurrent use is defined as 7 days of cumulative concurrent use in the previous 30 days prior to opioid dispensation based on days' supply". I am still not sure how concurrent use with benzodiazepine was defined in this study. In my understanding, concurrent use indicates concomitant use of opioids and benzodiazepines, but in this study benzodiazepines were identified within 30 days prior to opioid dispensation, which means that the opioid has not been started yet during the feature identification window. So how concurrent use of opioids and benzodiazepines occur within 30 days prior to opioid dispensation? Please clarify. 7. In the discussion, authors described that the current ML models can support PMPs and could be used to identify the opioid users at a high risk of developing subsequent adverse outcomes. Could authors add more discussions about how health department can utilize the ML models? Especially, the current model includes around 60 features (I am not sure the exact number as not clarified in the manuscript), which means that the users need to capture all the information to be able to make a prediction, a big burden for the users of this model. To make the model more practically useful, why don't authors rerun a decreased ML models with only the most important features shown in eFigure2 and eFigure 3 and assess the discrimination and calibration of the reduced model? To balance between the prediction power and use burden, I highly suggest authors rerunning some reduced ML models and see how the model will perform. 8. Suggest authors double checking the manuscript meticulously before next submission. There are many minor issues that deserve attention. For example: a. Page 4, line 35: Add a % after 13.38 b. Page 7, line 20: Please put DNN and GBM in the parenthesis instead of the full name of DNN and GBM c. Page 11, line 14: There was an extra "." after "training data" d. Some paragraphs had an indent, whereas others did not have. Please be consistent e. Please improve the resolution of all figures, especially Figure 1 and eFigure 3, which are difficult to read. f. Figure 2: Unify the plot size and make the 4 figures align with each other.

	g. Please make the first column in all the tables align on the left. For example, in table 1, please align the column of “Metric” on the left. h. eFigure 2: Some variables (e.g., Num_Fills_30, Doctor_Risk_30) are difficult to understand what they stand for. Please spell their full name. i. Font size of table 1 seems smaller than the main text. j. Font size of table 2 was not consistent. The part from “2019 Center for Medicare ……” had a larger font size than others. There are many other this type of minor issues in the manuscript that need attention. Therefore, I strongly suggest authors having a meticulous check before next submission.

REVIEWER	Deng, Hao Massachusetts General Hospital, Department of Anesthesia Critical Care and Pain Medicine
REVIEW RETURNED	25-Feb-2021

GENERAL COMMENTS	The authors properly addressed all of my comments except the one regarding missing data statement. It is interesting to see no missing data at all for health administrative databases, because it often contains large amount of missing data in practice. Is it possible for the authors to elaborate why there was no missing data? Is there no missing data even for the predictors. For instance, is there a specific regulation to mandate complete capture of all data elements? Besides this point, I think this work is publishable.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Dr. Lili Zhou, AbbVie Inc

Comments to the Author:
This manuscript entitled “Safe opioid prescribing: a machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada” aimed to predict a 30-day risk of hospitalization, emergency department visit, or mortality following an opioid prescription dispensation among the opioid users residing in Alberta, Canada, using machine learning techniques. This study used different types of machine learning models to maximize the prediction power. Overall, this is a relevant topic in the field given the high opioid prescribing in Canada. The authors have well addressed my comments on the study design and methods, but there are many notable minor concerns in the manuscript needed to be clarified in order to ensure the validity of the findings. Areas could be improved to strengthen the manuscript and study are listed below.

Response:

We will amend the manuscript according to Reviewer’s comments

1. (Page 5, line 10): The authors claimed that one strength of this study was “The study population is the entire provincial population and is generalizable to other populations in Canada and beyond”. Suggest removing this strength as it is not appropriate from reviewer’s perspective. This study includes the entire population living in Alberta, Canada, so it is confident to say that results are well

representative for this province. However, it is not appropriate to say that the results can be generalizable to other population in Canada because this study did not include participants from other regions of Canada and there are lack of evidence indicating that the opioid use pattern and subsequent risks were the same in other regions in Canada as in Alberta. Please also remove the relevant sentences in the discussion section.

Response: Vishal

The Reviewer is correct and we will amend as recommended:

- Bullet number 2 will be removed entirely and,
- Similar wording from Discussion will also be removed.
    - Removed: “However, our analyses were done on a large population and these results would be expected to be generalizable to the vast majority of patients.”

2. Could authors list all the ingredients (e.g., hydrocodone, oxycodone) included in opioids in this study in a supplemental table? Did authors also include opioid antitussives used for treatment of cough in the study? Did authors include both oral and parenteral opioids? The previous paper (citation number 10: Lo-Ciganic et al published in JAMA Open) that cited in this manuscript excluded individuals with only parenteral opioid prescription and/or cough or cold medications prescriptions containing opioids as these medications are either for acute pain relief or have low addictive potency. Suggest authors applying the same exclusion criteria in this study.

Response:

We will amend eTable 5 with the ATC codes used to identify the opioid molecules for this study and provide a reference for readers requiring more details of the ATC codes. We will add the following wording in the Methods section to clarify as recommended by the Reviewer:

“Anatomical Therapeutic Chemical classification (ATC) codes were used to identify opioid dispensations”

We included ALL opioid formulations in order to use all of the available data for ML training because provincial mandates require all dispensations to be uploaded to PIN, resulting in full capture of all opioids, including parenteral and anti-tussive formulations; we identified opioids using ATC codes, which is catalogued in PIN and are able to calculate daily OME’s for all opioid dispensations. With respect, we will not make these exclusions as recommended by the Reviewer as we believe our study’s exclusion criteria are sufficient.

3. (Page 9, line 37): Authors described the primary outcome of this study a composite of a DRUG-

RELATED hospitalization, ED visit or mortality within 30 days of an opioid dispensation. How did authors determine whether the hospitalization, ED visit or mortality was caused by opioid use? Please clarify it.

Response:

We used ICD-10 codes identified in the literature cited for opioid related events. Each hospitalization, emergency visit and death in DAD, NACRS and VS, respectively, use ICD-10 codes (up to 25) to identify causes. We used these fields in each data base to identify opioid related events with the specific ICD codes in eTable 2 and the Methods section of the manuscript.

We will clarify the Reviewer's point by adding the following wording to the Methods section:

"The primary outcome was a composite of a drug-related hospitalization, emergency department (ED) visit or mortality within 30 days of an opioid dispensation based on ICD-10 codes identified from DAD, NACRS and Vital Statistics (T40, F55, F10-19; eTable 2 in Supplement)"

4. (Page 9, line 47): Authors stated "To address this, we relied on specifying balanced class weightage for supporting algorithms". Could authors add more discussions of advantages of using this approach over the others that authors did not deem suitable. Also, please cite a reference here.

Response:

With regards to oversampling, we did not use this method because it introduces more bias into the training data set by generating new samples; on a side note, we ran an iteration using under sampling and its AUC was similar to the class weightage method.

We will add a reference as recommended by the Reviewer and clarify the wording as follows:

"Class weightage is a commonly used method to address class imbalance along with over and under-sampling approaches. However, oversampling, which involves generating new opioid dispensations from the original data distribution and prone to introducing bias, is difficult due to the categorical nature of the data and beyond the scope of this study. With under-sampling, which takes samples from the majority class (in this case no 30-day event after dispensation), we would not be able to use all of the information provided by the data in instances with no outcome. Hence, we decided to use the class weightage method which does not alter the data distribution. Instead, the learning process is adjusted in a way that increases the importance of the positive class (instances that led to a 30-day event)."

We will add these as references:

- https://link.springer.com/article/10.1186/s40537-019-0192-5
- King, G. and Zeng, L., 2001. Logistic regression in rare events data. Political analysis, 9(2), pp.137-163.

5. How many exact features were included in the model? I could not find the specific number in the

text. Please clarify it. In addition, in my previous comment, I asked why only Elixhauser comorbidities were included in the model? Shouldn't including more comorbidity features would further improve the model prediction power? In addition, for the feature category of healthcare utilization, only two features, including number of hospitalizations and number of unique providers, were included. Why not including more health utilization features to increase the prediction power (e.g., number of ED visits)? I understand that authors want to minimize the feature numbers by only including the potential strong predictors, but could authors clarify why only Elixhauser comorbidities and two healthcare utilization features were used? My concern is that authors may miss some important features. Thank you!

Response:

The number of features were 283 using all the databases (DAD, NACRS, PIN, VS). We will add wording to Results section and Table 3 to clarify according to Reviewer's recommendations:

"As described above, we categorized our candidate features into four groups (eTable 5 in Supplement). When using all of the databases, the total number of features was 283 and 34 when considering only co-morbidities (Table 3)."

Thanks for the comment. We tried to keep our models to data that would be routinely available by most health departments and PMPs. We used the Elixhauser list because it is a common and comprehensive co-morbidity measure in Canada, and elsewhere, in many health systems. Indeed, it is often used for hospital resource planning derived from hospitalization and billing records; it is a commonly used measure of co-morbidity in a variety of settings in Canada. However, we did add extra conditions (injury and poisonings) which are not in the Elixhauser index but are noted in the literature we referenced. We agree that if comorbidities are predictive it should increase the power. Beyond the 31 categories included in the Elixhauser index, plus the injury and poisonings, we are unclear which potential comorbidities could be informative in our models. Based on the literature, we have included all of the major comorbidities in our approach. However, if there are additional comorbidities the reviewer thinks we may have missed we are happy to further evaluate these to see if it assists in our models' prediction. We believe the Elixhauser list is a comprehensive measure of health for an individual.

Health care utilization: Thanks for the comment. This was an oversight on our part in terms of the presentation and wording in the Methods and Results. We actually did include number of emergency visits but it was included in a composite flag called "number of hospitalizations". We will clarify this issue as recommended by the Reviewer as follows:

- amend eTable 5 to include: "flags for previous hospitalizations and emergency department visits"
- amend Methods: "health care utilization (number of unique providers, number of hospital and emergency department visits),"

We tried to leverage the data sufficiently to predict our outcomes using these features.

6. Authors included a feature of “concurrent use with benzodiazepines”. Reviewer asked for the clarification of definition of this feature. In the response letter, authors said that “concurrent use is defined as 7 days of cumulative concurrent use in the previous 30 days prior to opioid dispensation based on days’ supply”. I am still not sure how concurrent use with benzodiazepine was defined in this study. In my understanding, concurrent use indicates concomitant use of opioids and benzodiazepines, but in this study benzodiazepines were identified within 30 days prior to opioid dispensation, which means that the opioid has not been started yet during the feature identification window. So how concurrent use of opioids and benzodiazepines occur within 30 days prior to opioid dispensation? Please clarify.

Response:

In this study, every opioid dispensation (not just the incident dispensation) was used as a potential instance to predict risk of 30-day events from. As a result, there may be active opioid prescriptions prior to the “reference” opioid dispensation that is being used for that 30-day prediction. Thus, concurrent use could have been present. For example, there might be a morphine dispensation for Patient A on Jan 1, Feb 1, and March 1. Although concurrency could not happen for the Jan 1 dispensation when looking at 30-day risk, there could be concurrency for the Feb 1 or March 1 dispensations of the opioid as concurrent use was flagged if there were other opioid and bzd dispensations within the 30 days prior to Feb 1, or March 1 dispensations.

We will clarify this point as recommended by Reviewer in the Methods section as follows:

“Every opioid dispensation, not just the incident one, was used as a potential instance to predict the risk of our outcome.”

7. In the discussion, authors described that the current ML models can support PMPs and could be used to identify the opioid users at a high risk of developing subsequent adverse outcomes. Could authors add more discussions about how health department can utilize the ML models? Especially, the current model includes around 60 features (I am not sure the exact number as not clarified in the manuscript), which means that the users need to capture all the information to be able to make a prediction, a big burden for the users of this model. To make the model more practically useful, why don’t authors rerun a decreased ML models with only the most important features shown in eFigure2 and eFigure 3 and assess the discrimination and calibration of the reduced model? To balance between the prediction power and use burden, I highly suggest authors rerunning some reduced ML models and see how the model will perform.

Response:

Our ML classifiers were trained on data that is commonly available to PMPs and other health departments. However, the Reviewer makes a good suggestion and we will include a table showcasing the discrimination performance of the ML models with sub-groups of the available data which all PMPs should easily have access to in some capacity (Table 3). However, like any system, if PMP’s are interested in deploying this kind of risk assessment system according to their own data availability/accessvalidity testing and possibly feature reduction within their data source will be required. ML techniques have to be modified to the environment they are deployed and this is why it is so important to be transparent on what features are being used – as we have in our paper. All ML models can be tailored according to their needs and capabilities as defined by data access, which is clearly defined in the ML process that we referenced.

We will run the XGBoost classifier using the most important features identified in eFigure 3. However, health departments and PMPs are limited by data access and not variable importance; health departments cannot choose features based on importance from our ML models because they are limited by data access and availability.

The Reviewer makes a valid point and we will amend the Discussion as follows:

"This study suggests that ML risk prediction can support PMPs, especially if readily available administrative health data is used. PMPs currently use population-based guidelines which we, and others, have shown cannot predict absolute individual risk. The ML process allows for flexibility in model training, validation and deployment to specific settings and data in which, for the case of PMPs, high risk patients can be identified and targeted for intervention either at the patient or provider level. For example, a ML classifier can be trained on accessible data to create an aggregated list of "high risk" patients at regular time intervals to identify points of intervention. Moreover, ML classifiers can be retrained over time as changes in populations and trends in prescribing occur and are therefore specific to the population unlike broadly based guidelines. Further research can assess whether implementation of a ML-based monitoring system by PMPs leads to improved clinical outcomes within their own jurisdictions and whether other available features or feature reduction can yield sufficiently valid results for their own intended purposes."

As recommended by Reviewer, we will rerun the XGBoost model using the features identified in eFigure 3 and report the AUC in the eFigure 3 eSupplement. We will also clarify variable names in the eFigure 3 to make it clearer for the reader. eFigure 3 was based on XGBoost only.

We added this to eFigure 3A

"Note: Training and validating the XGBoost classifier with these features alone resulted in an AUC of 0.877 in the 2018 validation set."

8. Suggest authors double checking the manuscript meticulously before next submission. There are many minor issues that deserve attention. For example:

a. Page 4, line 35: Add a % after 13.38

b. Page 7, line 20: Please put DNN and GBM in the parenthesis instead of the full name of DNN and GBM

c. Page 11, line 14: There was an extra "." after "training data"

d. Some paragraphs had an indent, whereas others did not have. Please be consistent

e. Please improve the resolution of all figures, especially Figure 1 and eFigure 3, which are difficult to read.

f. Figure 2: Unify the plot size and make the 4 figures align with each other.

g. Please make the first column in all the tables align on the left. For example, in table 1, please align the column of "Metric" on the left.

h. eFigure 2: Some variables (e.g., Num_Fills_30, Doctor_Risk_30) are difficult to understand what they stand for. Please spell their full name.

i. Font size of table 1 seems smaller than the main text.

j. Font size of table 2 was not consistent. The part from "2019 Center for Medicare ……" had a larger font size than others.

There are many other this type of minor issues in the manuscript that need attention. Therefore, I strongly suggest authors having a meticulous check before next submission.

Response:

All edits were made as recommended by Reviewer and manuscript was thoroughly re-checked for formatting issues.

Reviewer: 2
Dr. Hao Deng, Massachusetts General Hospital

Comments to the Author:
The authors properly addressed all of my comments except the one regarding missing data statement. It is interesting to see no missing data at all for health administrative databases, because it often contains large amount of missing data in practice. Is it possible for the authors to elaborate why there was no missing data? Is there no missing data even for the predictors. For instance, is there a specific regulation to mandate complete capture of all data elements? Besides this point, I think this work is publishable.

Response:

Reviewer is correct in that there are reporting mandates which require reporting of all hospitalizations, emergency visits, provider visits, and prescription dispensations and details for each instance. Thus, no missing data. All predictors came from databases with these mandates.

VERSION 3 – REVIEW

REVIEWER	Zhou, Lili AbbVie Inc, Health economics and outcome research
REVIEW RETURNED	24-Apr-2021

GENERAL COMMENTS	Comments to the Authors This manuscript entitled “Safe opioid prescribing: a machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada” aimed to predict a 30-day risk of hospitalization, emergency department visit, or mortality following an opioid prescription dispensation among the opioid users residing in Alberta, Canada, using machine learning techniques. This study used different types of machine learning models to maximize the prediction power. Overall, this is a relevant topic in the field given the high opioid prescribing in Canada. The authors have well addressed my comments on the study design and methods, but there are several other notable minor concerns in the manuscript needed to be clarified in order to ensure the validity of the findings. Areas could be improved to strengthen the manuscript and study are listed below. 1. (Page 4, line 13): It stated that “all patients over 18 years of age”, whereas in the methods section on page 8, line 23, the authors described the eligibility of age as “18 years of age and

	older". Please be consistent in the age criterion applied in this study. 2. (Page 7, line 32): The authors noted that the previous study "did not emphasize other performance metrics required to assess clinical utility". Please add several performance metrics in a parenthesis that authors think are important to examine. 3. (Page 8, line 50-51): It mentioned that "ATC codes were used to identify opioid dispensations (eSupplement)". I think the authors refer to eTable 5 here. Please clarify it in the text. This sentence was added to address my previous comments regarding what specific drugs (e.g., hydrocodone, oxycodone) were considered as opioids in this study. I think that the authors misunderstood my comments. So, I'd like to clarify it here. Please list the specific drug names rather than the ATC codes used to identify these ingredients. From my perspective, the specific drugs counted as opioids are more of interest. 4. (Page 9, line 26-31): Please move the sentence "All analyses were done using ……" to the end of the section with the heading of "Statistical Analyses and Machine-Learning Prediction Evaluation". It is not appropriate to put this sentence in the current spot. 5. (Page 9, line 49): The authors described the ICD10 codes (T40.x, F55.x, F11.x-F19.x) used to identify opioid related hospitalization, ED visit, and mortality. However, those ICD10 codes can also indicate the drug overdose, abuse, and use disorder associated with many other psychoactive drugs. Opioids can only account for a portion of all the cases. How authors are sure that these cases suggest opioid-related adverse events? Please clarify. 6. (Page 12, line 41-46): The authors stated using the 2019 CMS opioid safety measure and 2017 Canadian opioid prescribing guideline to assess its prediction power. Could authors provide more information of how the ML models were employed using these guidelines? For example, the 2019 CMS opioid guideline required the opioid use evaluated in a 12-month observation. In this study, did the authors use the whole 2017 data to identify those predictors (e.g., high dose, concurrent use with benzodiazepines) suggested from the guideline and then predict the opioid-related adverse events in the first 30 days of 2018? There is little information available in terms of how these guidelines were applied in assessing the prediction performance. Please add more details describing the methods. 7. All the figures in this manuscript have a low resolution. The texts in the figures look blurry. Please improve the resolution to meet the publication requirements. Also, if there are multiple graphs in a figure (e.g., Figure 2). Please make sure the size of all the graphs the same and align with each other. 8. (Table 3, eTable 3, eTable 4): Please add a footnote in Table 3 spelling out the abbreviations presented in the table. For eTable3 and eTable 4, please be consistent in the formatting of the variable of age.

REVIEWER	Deng, Hao Massachusetts General Hospital, Department of Anesthesia Critical Care and Pain Medicine
REVIEW RETURNED	20-Apr-2021

GENERAL COMMENTS	I have no further comments.

VERSION 3 – AUTHOR RESPONSE

Response to Reviewer: 1
Dr. Lili Zhou, AbbVie Inc

This manuscript entitled “Safe opioid prescribing: a machine learning approach to predicting 30-day risk after an opioid dispensation in Alberta, Canada” aimed to predict a 30-day risk of hospitalization, emergency department visit, or mortality following an opioid prescription dispensation among the opioid users residing in Alberta, Canada, using machine learning techniques. This study used different types of machine learning models to maximize the prediction power. Overall, this is a relevant topic in the field given the high opioid prescribing in Canada. The authors have well addressed my comments on the study design and methods, but there are several other notable minor concerns in the manuscript needed to be clarified in order to ensure the validity of the findings. Areas could be improved to strengthen the manuscript and study are listed below.
Response:
Thank you for the comments. We will amend the manuscript according to Reviewer comments.

1. (Page 4, line 13): It stated that “all patients over 18 years of age”, whereas in the methods section on page 8, line 23, the authors described the eligibility of age as “18 years of age and older”. Please be consistent in the age criterion applied in this study.

Response:
Reviewer is correct. We will amend the Abstract so that the wording is consistent with the Methods as follows:
“Participants included all patients 18 years of age and older who received at least one opioid dispensation”

2. (Page 7, line 32): The authors noted that the previous study “did not emphasize other performance metrics required to assess clinical utility”. Please add several performance metrics in a parenthesis that authors think are important to examine.

Response:
Reviewer makes a valid recommendation. We weill amend this section as follows:
“This study relied on c-statistics to evaluate their ML models and did not emphasize other performance metrics (e.g., positive likelihood ratios, pre and post-test probabilities) required to assess clinical utility that are recommended by medical reporting guidelines[11,13,19,20].”

3. (Page 8, line 50-51): It mentioned that “ATC codes were used to identify opioid dispensations (eSupplement)”. I think the authors refer to eTable 5 here. Please clarify it in the text. This sentence

was added to address my previous comments regarding what specific drugs (e.g., hydrocodone, oxycodone) were considered as opioids in this study. I think that the authors misunderstood my comments. So, I'd like to clarify it here. Please list the specific drug names rather than the ATC codes used to identify these ingredients. From my perspective, the specific drugs counted as opioids are more of interest.

Response:
Reviewer is correct. We will amend the Methods section as follows:
"eTable 5" and remove "eSupplement".

Sorry for the confusion. We will also amend the manuscript to include names of opioid molecules used in this study. We will add the names to eTable 5 and include wording in the Methods section as follows:
eTable 5:

Opioid molecules used in this study	alfentanil, butorphanol, codeine, diamorphine, fentanyl, hydrocodone, hydromorphone, meperidine, morphine, oxycodone, oxymorphone, pentazocine, sufentanil, tapentadol, tramadol

Methods:
"Anatomical Therapeutic Chemical classification (ATC) codes[22] were used to identify opioid dispensations and their respective opioid molecules (eTable 5)"

4. (Page 9, line 26-31): Please move the sentence "All analyses were done using ……" to the end of the section with the heading of "Statistical Analyses and Machine-Learning Prediction Evaluation". It is not appropriate to put this sentence in the current spot.

Response:
Reviewer makes a valid point. We will move this sentence as described by Reviewer.

5. (Page 9, line 49): The authors described the ICD10 codes (T40.x, F55.x, F11.x-F19.x) used to identify opioid related hospitalization, ED visit, and mortality. However, those ICD10 codes can also indicate the drug overdose, abuse, and use disorder associated with many other psychoactive drugs. Opioids can only account for a portion of all the cases. How authors are sure that these cases suggest opioid-related adverse events? Please clarify.

- Opioid related hospitalizations are coded in a variety of ways, we used these to be all inclusive… https://www.bmj.com/content/362/bmj.k3207
-

Response:
Reviewer makes a valid observation. We used these codes because opioid related events can be coded in a variety of ways and because other researchers have included them in their research which was published in BMJ (Gomes paper at https://www.bmj.com/content/362/bmj.k3207); we did this to be on the cautious side to be as inclusive as possible so we don't miss any events. Moreover, opioids are always cited as a potential contributing factor, irrespective of the actual agent that led to the hospitalization, when any overdose occurs and is consistent with how the opioid epidemic is being tracked by public health agencies across North America. Nevertheless, we will change the wording in the Measures and Outcomes section to clarify the Reviewer's point and emphasize the reference mentioned above as follows:

"The primary outcome was a composite of a drug-related hospitalization, emergency department (ED) visit or mortality within 30 days of an opioid dispensation based on ICD-10 codes used by others and identified from DAD, NACRS and Vital Statistics (T40, F55, F10-19; eTable 2 in Supplement)[2,10,30]"

6. (Page 12, line 41-46): The authors stated using the 2019 CMS opioid safety measure and 2017 Canadian opioid prescribing guideline to assess its prediction power. Could authors provide more information of how the ML models were employed using these guidelines? For example, the 2019 CMS opioid guideline required the opioid use evaluated in a 12-month observation. In this study, did the authors use the whole 2017 data to identify those predictors (e.g., high dose, concurrent use with benzodiazepines) suggested from the guideline and then predict the opioid-related adverse events in the first 30 days of 2018? There is little information available in terms of how these guidelines were applied in assessing the prediction performance. Please add more details describing the methods.
Response:
Reviewer makes a valid observation.

- CMS and CAD guidelines are not ML models; we used the guidelines as rules to predict the 30-day risk of event at the time of dispensation for the entire 2018 validation set so that they can be directly compared to the ML models (same time period as ML model). The predictors for the guidelines are the rules we used when coding. We will clarify this in the Methods section and in the caption in Table 2 as follows:

Methods:
"This was done by using the guidelines as "rules" when coding for the 30-day risk of event at the time of each opioid dispensation on the entire 2018 validation set."
Table 2:
"These guidelines were used as rules to predict the 30-day risk of event at the time of opioid dispensation."

- As suggested by guidelines, we used their specified time periods (CMS only had timelines, CAD guidelines have no timeline)
- For CMS, the assessment period is 180 days prior; we used only 2018 validation data for this evaluation

The timelines expressed in the guidelines are somewhat open to interpretation and could be constrained by data availability. Nevertheless, we will clarify the Reviewer's point and amend the footnote under Table 2 as follows:
"**The CMS guidelines specify 90 or more days at >120 OME and concurrent use of opioids and benzodiazepines for 30 days or more within an assessment period of 180 days."

7. All the figures in this manuscript have a low resolution. The texts in the figures look blurry. Please improve the resolution to meet the publication requirements. Also, if there are multiple graphs in a figure (e.g., Figure 2)Please make sure the size of all the graphs the same and align with each other.
Response:
Reviewer is correct. We will do our best to improve the resolution. We noticed that the conversion from word.doc to pdf results in some loss of resolution. eFigure 3C is the most troublesome, we have improved it and hopefully it will show better upon conversion to pdf (it does on our end). Nevertheless, if it does not, we can remove this figure from the eSupplement if the Editor wishes. We will let the Editor decide.
We provided 1200 dpi image resolution in word docs.

As well, we will provide the word doc files to the editor and perhaps they can offer an opinion and other options. With the corrections to resolution we made, zooming in the pdf file greatly improves resolution if needed.
Figures 1,2,3,4 have much better resolution now in pdf.

eFigure 3C and 4 are much improved as well.

All figures now are in better resolution and are clearer on our end.

8. (Table 3, eTable 3, eTable 4): Please add a footnote in Table 3 spelling out the abbreviations presented in the table. For eTable3 and eTable 4, please be consistent in the formatting of the variable of age.
Response:
We will make the changes as specified by the Reviewer.
For Table 3:
"PIN- Pharmaceutical Information Network; DAD- Discharge Abstract Database; NACRS- National Ambulatory Care Reporting System"

eTable 3 and 4:
Formats were aligned and are now consistent.