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)	Develop an ADR prediction system of Chinese herbal injections containing Panax notoginseng saponin: a nested case-control study using machine learning
AUTHORS	Wu, Xing-Wei; Zhang, Jia-Ying; Chang, Huan; Song, Xue-Wu; Wen, Ya-Lin; Long, En-Wu; Tong, Rong-Sheng

VERSION 1 – REVIEW

REVIEWER	Xuan, Jianwei Sun Yat-Sen University
REVIEW RETURNED	25-Feb-2022

GENERAL COMMENTS	The manuscript focused on a critical area of machine leaning application to support ADR management. The development of this area interests will eventually significantly help proactive clinical management of potential ADR associated with various medicines. Would support publish this manuscrip. One minor suggestion to author is that to find a native english speaker to poolish the writing better.

REVIEWER	Chen, Yongchuan Third Military Medical University, Department of Pharmacy
REVIEW RETURNED	05-Mar-2022

GENERAL COMMENTS	nteresting study about to develop an adverse drug reactions (ADR) antecedent prediction system using machine learning algorithms to provide the reference for security usage of Chinese herbal injections containing Panax notoginseng saponin in clinical practice. The study is well designed, and the manuscript is generally well structured. The authors used eighteen machine learning algorithms to establish 1080 ADR prediction models. An ADR prediction system with Chinese herbal injections containing Panax notoginseng saponin developed by the best model had high accuracy and precision, and had potential value for clinical application, which is innovative and with certain clinical value. Of course, the limitations of this study are due to the retrospecive analysis, so more data were needed to further evaluate the model prediction performance. For all that, I hope the authors to clatify the following questions. The data collection of the method part is not particularly clear, in the begin the author said some patient's data come from the National Center for ADR Monitoring, but in the end, the author said all patient's data were from 5 medical institutions in Sichuan Province? Please descript it more accurate.

	The author divided the data into a training set and a test set according to 8:2, Please explain the rationality of this proportion distribution. A series of data preprocessing methods were adopted in the study, however, the overview of the dataset before and after data clean is not clear, please add relevant clarification. In sample size assessment section, the std of AUC of ROC was not stable according to the figure, is that acceptable? What are the possible reasons?

REVIEWER	Do, Quan Mayo Clinic Rochester
REVIEW RETURNED	03-Jun-2022

GENERAL COMMENTS	This is quite an interesting topic. However, maybe the language barrier caused many limitations to this paper. For study design, the description of the study's participants was not clear. "Participants: All patients were from 5 medical institutions in Sichuan Province from 30 January 2010 to December 2018" Did you mean all patients who were admitted during that period were automatically enrolled into this study? Were they informed about this study and agreed to participate? During this period (2018), how many people had multiple admissions and how did you deal with multiple admissions and so on? In addition, if the authors described the selection method clearer, it would be more informative for the readers: "A total of 530 patients were enrolled in this study, of which 106 patients had ADR. ADR patients included 50 (47.17%) males and 56 (52.83%) females." Also, it might be better if the authors described the study population with more details Regarding the methods used for this study, please describe with more details when the techniques used may lead to bias: 1. How did you deal with multiple lab results? For example, for each admission, there are multiple lab results for each lab test. How did you deal with this issue? 2. Missing Data (Data filling): how many missing data (in percentage)? Why you decide to use Random Forest method for replacing the missing data? 3. Variable selection: Which variables were selected, and which variables were dropped after this step? Was there any difference result from the 2 methods used? 4. why did you use 18 algorithms? Can you group these algorithms into several groups then select the most representative algorithms in each group? What are the advantages and disadvantages of these algorithms? English issues: Proofreading is needed. There are many sentences that were not understandable. For example, "The data was standardized and divided into a training set and a test set according to 8:2" Did you mean, the data was randomly split into training set and testing set by the ratio of 8:2?

	Because of the above sentence, I can surmise that you mentioned a ratio of 1:4 in the following sentence, but I didn't get the idea of which was 1 and which was 4: "A nested case-control study was used to randomly match patients without ADR who using Chinese herbal injections containing Panax notoginseng saponin from the EMR system according to 1:4"

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Dr. Jianwei Xuan, Sun Yat-Sen University

Comment 1: *The manuscript focused on a critical area of machine leaning application to support ADR management. The development of this area interests will eventually significantly help proactive clinical management of potential ADR associated with various medicines. Would support publish this manuscript. One minor suggestion to author is that to find a native English speaker to polish the writing better.*

Response: Thank you for your suggestion. We have polished the manuscript.

Reviewer: 2

Dr. Yongchuan Chen, Third Military Medical University

Comment 1: *The data collection of the method part is not particularly clear, in the begin the author said some patient's data come from the National Center for ADR Monitoring, but in the end, the author said all patient's data were from 5 medical institutions in Sichuan Province? Please descript it more accurate.*

Response: We feel sorry for the data collection in the method part does not clear. ADR patients who used Chinese herbal injections containing Panax notoginseng included in this study were from the National Center for Adverse Drug Reaction Monitoring reported by 5 hospitals in Sichuan Province from January 2010 to December 2018. Then, a nested case-control study was used to randomly match patients without ADR from the EMR system of the 5 medical institutions. The ratio of patients with ADR to those without ADR was 1:4. And this section has been modified in the method.

Comment 2: *The author divided the data into a training set and a test set according to 8:2, Please explain the rationality of this proportion distribution.*

Response: Thanks for your questions. The common ratio of the training set and test set are 7:3, 8:2, and 9:1. A total of 530 patients were included in this study. When the training set and test set are divided according to 7:3, the small size of the training set will lead to insufficient model training. But if divided according to 9:1, the small size of the test set makes it difficult to accurately evaluate the

predictive performance of the model. Therefore, we divided the data into a training set and a test set according to 8:2.

Comment 3: *A series of data preprocessing methods were adopted in the study, however, the overview of the dataset before and after data clean is not clear, please add relevant clarification.*

Response: Thank you for your suggestions. A total of 530 patients and 83 variables were included in this study. In the column deletion, 20 variables (missing data >90%, or a single category >90%, or the coefficient of variation <0.1) were deleted. And 63 variables were included in the following study. In addition, there were 1,290 (3.86%) missing values, which were replaced by 4 data filling methods. We used Lasso or Boruta for variable selection, and the results of the two variable selection methods were shown in the table 1. This section has been added in the Supplementary materials.

Table 1 Results of different variable preprocessing methods

Method	Included variables
Column deletion	X1, X2, X3, X5, X7, X8, X12, X13, X14, X15, X16, X17, X18, X19, X20, X21, X22, X28, X29, X30, X31, X32, X33, X34, X35, X36, X39, X40, X41, X42, X43, X44, X45, X46, X51, X52, X54, X55, X56, X57, X58, X59, X60, X61, X62, X63, X65, X66, X67, X68, X71, X72, X73, X74, X75, X76, X77, X78, X79, X80, X81, X82, X83
Lasso	X1, X2, X18, X29, X30, X31, X33, X51, X52, X54, X55, X65, X66, X68, X78
Boruta	X1, X2, X5, X12, X13, X16, X17, X18, X20, X29, X30, X31, X33, X39, X40, X51, X52, X54, X55, X63, X66, X67, X68, X74, X75, X77, X78, X79

Variable names were shown in Supplementary Table 2.

Comment 4: *In sample size assessment section, the std of AUC of ROC was not stable according to the figure, is that acceptable? What are the possible reasons?*

Response: Thanks for your questions. The results of the sample size assessment (Figure 4) showed that the std of AUC of ROC was not stable when the sample size was between 10% and 30%. This is due to the insufficient size of the training set made the predictive performance of the model fluctuate significantly. With the continuously increased size of sample data, the AUC values continued to increase, the predictive performance was stabilized, and the std of AUC of ROC gradually decreased.

Reviewer: 3

Dr. Quan Do, Mayo Clinic Rochester

Comment 1: *For study design, the description of the study's participants was not clear.*

"Participants: All patients were from 5 medical institutions in Sichuan Province from

30 January 2010 to December 2018"

Did you mean all patients who were admitted during that period were automatically enrolled into this study? Were they informed about this study and agreed to participate?

During this period (2018), how many people had multiple admissions and how did you deal with multiple admissions and so on?

Response: We feel sorry for the expression not being clear in this part. ADR patients included in this study were from the National Center for Adverse Drug Reaction Monitoring reported by 5 medical institutions in Sichuan Province from January 2010 to December 2018. Then, we used a nested case-control study in which patients without ADR were enrolled from the 5 hospitals. The ratio of patients with ADR to those without ADR was 1:4. This study was approved by the Ethics Committee of Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital. Due to the retrospective nature of the study, informed consent was waived. And we hid the patients' personal information during the study. This section has been modified in the method. In addition, all patients were included according to their first admission.

Comment 2: *In addition, if the authors described the selection method clearer, it would be more informative for the readers:*

"A total of 530 patients were enrolled in this study, of which 106 patients had ADR. ADR patients included 50 (47.17%) males and 56 (52.83%) females."

Also, it might be better if the authors described the study population with more details

Response: Thank you for your suggestion. The demographic and clinical characteristics of the patients were shown in table 1. And this section has been added in the Supplementary materials.

Table 1 Demographic and clinical characteristics of the patients

Variables	Number
Gender	
Male	250(47.17)
Female	280(52.83)
Age (years)	
≤ 44	121(22.83)
45 ≤ Age ≤ 59	193(36.42)
60 ≤ Age ≤ 74	132(24.91)
≥ 75	84 (15.85)
Body mass index (BMI, kg/m^2)	

< 18.5	48(9.06)
18.5 ≤ BMI ≤ 23.9	275(51.89)
≥ 24	175(33.02)
Charlson comorbidity index (Score)	
0	104(19.62)
1 or 2	190(35.85)
3 or 4	123(23.21)
≥ 5	113(21.32)

Data presented as number (%)

Comment 3: *How did you deal with multiple lab results? For example, for each admission, there are multiple lab results for each lab test. How did you deal with this issue?*

Response: Thanks for your questions. For multiple lab results, in order to facilitate clinical application, we selected the last results of patients before the usage of medication. And this section has been added in the method.

Comment 4: *Missing Data (Data filling): how many missing data (in percentage)? Why you decide to use Random Forest method for replacing the missing data?*

Response: Thanks for your questions. After the column deletion (Variables with missing data >90%, or a single category >90%, or the coefficient of variation (CV) <0.1), there were 63 variables included in the following study. A total of 33,390 data were from 530 patients, including 1,290 (3.86%) missing values. We used the random forest method for replacing the missing data. Because the random forest method can better simulate the distribution of the original data than using the median or mode methods to replace the missing data directly.

Comment 5: *Variable selection: Which variables were selected, and which variables were dropped after this step? Was there any difference result from the 2 methods used?*

Response: Thanks for your questions. The results of the two variable selection methods were shown in table 2. The variable selected by Lasso was less than by Boruta, this was because Boruta selects all variables related to model performance, while Lasso removes more variables by constructing a penalty function. And this section has been added in the Supplementary materials.

Table 2 The results of the two variable selection methods

Method	Variables
Lasso	X1, X2, X18, X29, X30, X31, X33, X51, X52, X54, X55, X65, X66, X68, X78

Boruta	X1, X2, X5, X12, X13, X16, X17, X18, X20, X29, X30, X31, X33, X39, X40, X51, X52, X54, X55, X63, X66, X67, X68, X74, X75, X77, X78, X79

X1, Gender; X2, Age; X5, Genetic family history; X12, Blood pressure; X13, Charlson comorbidity index; X16, Respiratory diseases; X17, Nervous diseases; X18, Digestive diseases; X20, Orthopedic diseases; X29, Dose; X30, Anti-infective agents; X31, Cardiovascular medicines; X33, Respiratory medicines; X39, Medicines for hematopathy; X40, Endocrine agents or hormone drugs; X51, Dermatology medication; X52, Other traditional Chinese medicines or Chinese patent medicines; X54, Serum creatinine; X55, Renal function; X63, Albumin; X65, Globulin; X66, Albumin/globulin (A/G); X67, Aspartate aminotransferase; X68, Alanine aminotransferase; X74, Neutrophil granulocyte; X75, Lymphocyte percentage; X77, Eosinophils; X78, Red blood cell; X79, Hemoglobin.

Comment 6: *why did you use 18 algorithms? Can you group these algorithms into several groups then select the most representative algorithms in each group? What are the advantages and disadvantages of these algorithms?*

Response: Thanks for your questions. The 18 machine learning algorithms used in this study could be divided into the following categories, tree algorithms (Decision Tree, Extra Tree); Bayes algorithms (Bernoulli Naïve Bayes, Gaussian Naïve Bayes, Multinomial Naïve Bayes); ensemble algorithms (Bagging, Random Forest, Gradient Boosting, AdaBoost, eXtreme Gradient Boosting, Ensemble Learning); distance algorithms (K-Nearest Neighbor); and other types of algorithms (Latent Dirichlet Allocation, Logistic Regression, Passive Aggressive, Quadratic Discriminant Analysis, Stochastic Gradient Descent, Support Vector Machine). KNN and SVM algorithms have stable predictive performance. The ensemble algorithm has powerful data processing capability. eXtreme Gradient Boosting has a wide application in medical. And the tree algorithms are more interpretable. However, there are differences in the prediction performance of different machine learning algorithms. Therefore, our study selected the best performance model by comparing 18 machine learning algorithms.

Comment 7: *"The data was standardized and divided into a training set and a test set according to 8:2"*

Did you mean, the data was randomly split into training set and testing set by the ratio of 8:2?

Because of the above sentence, I can surmise that you mentioned a ratio of 1:4 in the following sentence, but I didn't get the idea of which was 1 and which was 4:

"A nested case-control study was used to randomly match patients without ADR who using Chinese herbal injections containing Panax notoginseng saponin from the EMR system according to 1:4"

Response: We feel sorry for the expression in this part does not clear.

In the data collection, we enrolled ADR patients from the National Center for Adverse Drug Reaction Monitoring reported by 5 hospitals in Sichuan Province. Then, a nested case-control study was used to randomly match patients without ADR from the EMR system of the 5 medical institutions. The ratio of patients with ADR to those without ADR was 1:4.

In the model establishment, the data were randomly split into a training set and a test set by the ratio of 8:2. The training set was used to build models, and the test set was used to evaluate the predictive performance of the models. And this section has been modified in the method.

VERSION 2 – REVIEW

REVIEWER	Chen, Yongchuan Third Military Medical University, Department of Pharmacy
REVIEW RETURNED	26-Jul-2022

GENERAL COMMENTS	The study is well designed, and the manuscript is generally well structured. The authors used eighteen machine learning algorithms to establish 1080 ADR prediction models, which had potential value for clinical application. In my opinion this article has reached the publishing level.

REVIEWER	Do, Quan Mayo Clinic Rochester
REVIEW RETURNED	01-Aug-2022

GENERAL COMMENTS	Thank you for working on the improvement of this paper. Most of my comments were resolved.