BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (http://bmjopen.bmj.com).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

BMJ Open

Developing an ADR prediction system of Chinese herbal injections containing Panax notoginseng saponin: a nested case-control study using machine learning

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-061457
Article Type:	Original research
Date Submitted by the Author:	27-Jan-2022
Complete List of Authors:	Wu, Xing-Wei; Sichuan Provincial People's Hospital Zhang, Jia-Ying ; Chengdu First People's Hospital Chang, Huan ; Sichuan Provincial People's Hospital Song, Xue-Wu ; Sichuan Provincial People's Hospital Wen, Ya-Lin ; Sichuan Provincial People's Hospital Long, En-Wu ; Sichuan Provincial People's Hospital Tong, Rong-Sheng; Sichuan Provincial People's Hospital,
Keywords:	Adverse events < THERAPEUTICS, Herbal medicine < THERAPEUTICS, Toxicity < THERAPEUTICS

SCHOLARONE™
Manuscripts

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

For peer review only

Develop an ADR prediction system of Chinese herbal injections containing Panax notoginseng saponin: a nested case-control study using machine learning

Xing-Wei Wu[1,2], Jia-Ying Zhang[3], Huan Chang[1], Xue-Wu Song[1,2], Ya-Lin Wen[1], En-Wu Long[1,2], Rong-Sheng Tong[1,2]

[1]Personalized Drug Therapy Key Laboratory of Sichuan Province, School of Medicine, University of Electronic Science and Technology of China, Chengdu 610072, China,

[2]Department of Pharmacy, Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital, Chengdu 610072, China,

[3]Department of Pharmacy, Chengdu First People's Hospital, Chengdu 610095, China

Correspondence to

Dr Rong-Sheng Tong, Personalized Drug Therapy Key Laboratory of Sichuan Province, School of Medicine, University of Electronic Science and Technology of China, Chengdu 610072, China. Department of Pharmacy, Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital, Chengdu 610072, China. E-mail: 318004031@qq.com

Word count: 2225

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Develop an ADR prediction system of Chinese herbal injection containing Panax notoginseng saponin: a nested case-control study using machine learning

ABSTRACT

Objective This study aimed 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.

Design A nested case-control study.

Setting National Center for ADR Monitoring and the Electronic Medical Record (EMR) system.

Participants All patients were from 5 medical institutions in Sichuan Province from January 2010 to December 2018.

Main outcomes/measures Information of patients with ADR who using Chinese herbal injections containing Panax notoginseng saponin was collected from the National Center for ADR Monitoring. A nested case-control study was used to randomly match patients without ADR from the EMR system according to 1:4. Eighteen machine learning algorithms were applied for the development of ADR prediction models. Area under curve (AUC), accuracy, precision, recall rate and F1 value were used to evaluate the predictive performance of the model. An ADR prediction system were established by the optimal model selected from the 1080 models.

Results A total of 530 patients from 5 medical institutions were included, and 1080

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

ADR prediction models were developed. Among these models, the AUC of the best capable one was 0.9141 and the accuracy was 0.8947. According to the parameters of the best model, a prediction system for the ADR of Panax notoginseng saponin has been established, which can realize the output of patient ADR risk.

Conclusion The prediction system developed based on the machine learning model in this study had good predictive performance and potential clinical application.

Key words Adverse drug reactions, Chinese herbal injection, Machine learning, Prediction system, Panax notoginseng saponin

Strengths and limitations of this study

- We first used machine learning to predict the ADR of Chinese herbal injection containing Panax notoginseng saponin.
- Eighteen machine learning algorithms were used 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.
- More than 80 factors including the patient's pathophysiological characteristics, clinical laboratory results, and medication conditions, were incorporated in our study.
- More data were needed to further evaluate the model prediction performance.

INTRODUCTION

Panax notoginseng saponins, as the main ingredients of Panax notoginseng (Buck.) F.H.Chen, has been widely used in the disease therapy of nervous system and cardio-

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

cerebral vascular system [1-4]. High frequency of adverse drug reactions (ADR) in Chinese herbal containing Panax notoginseng saponin has received widespread attention. Of all the adverse reactions, about 69.57% were caused by injections, mainly manifested as drug eruption (50.5%), allergic reaction (20.4%) and anaphylactic shock (9.7%), which can be life-threatening in severe cases [5].

At present, ADR is mainly monitored by spontaneous reporting system, case-control study, cohort study, prescription event monitoring and centralized hospital monitoring system. However, most of these methods have obvious hysteresis. Therefore, there is an increasing need to develop an ADR antecedent prediction system to prevent and avoid the occurrence of ADR in Chinese herbal injections containing Panax notoginseng saponin.

Machine learning, the core technology of artificial intelligence, is commonly used to build prediction models. In recent years, some prediction models for ADR have been established [6-10]. Based on a clustering method for the postprocessing of association rules, Lai et al. [6] developed an application of stepwise association rule mining to identify the associations between vaccine and multiple adverse events. In addition, Imai et al. [10] used artificial neural networks to evaluate vancomycin-induced nephrotoxicity. However, small sample size, incomplete patient information, and unsatisfactory predictive performance restrict the application of ADR prediction models in clinical practice. In view of these challenges, this study collected patients information in the National Center for ADR Monitoring and the Electronic Medical Record (EMR) system by a nested case-control study to establish an ADR prediction model of Chinese herbal

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

injections containing Panax notoginseng saponin, and develop an ADR prediction system based on machine learning algorithms to provide reference for clinical ADR management and prevention.

METHODS

Data collection

Information of patients with ADR who using Chinese herbal injections containing Panax notoginseng saponin from the National Center for ADR Monitoring was collected. 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. All patients were from 5 medical institutions in Sichuan Province from January 2010 to December 2018. This study was approved by the Ethics Committee of Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital.

Data cleaning

Variable assignment

Binary-state variables were directly assigned values of 0 or 1. According to whether in the normal range, clinical laboratory variables were assigned values of 1, 2 and 3 (1, below the normal range; 2, within the normal range; and 3, above the normal range).

Column deletion

Variables with missing data >90%, or a single category >90%, or the coefficient of variation (CV) <0.1 were deleted.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Data filling

There are 4 ways to data filling. No filling means to retain the original data directly. Simple filling refers to use the mean fill for continuous variables, the mode for disordered categorical variables, and the median for ordered categorical variables. Random Forest (RF) filling orders the column according to the number of missing data, and then the missing data was predicted and filled by RF model. RF improve filling refers to predict and fill the column with the least missing data, which was used as the input for the prediction and filling of other missing data.

Data sampling

No sampling: directly input the original data into the model. Random over sampler: random replication of data with fewer types to make the sample sizes of different types consistent, while random under sampler is to randomly delete data with more types. Synthetic minority oversampling technique (SMOTE) over sampler: synthesize new data by analyzing a small amount of original data. Borderline SMOTE over sampler: synthesize new data from borderline data.

Variable selection

No variable selection or use Lasso or Boruta for variable selection.

Model establishment

Through different data filling, data sampling and variable selection, 60 data sets were obtained. Eighteen machine learning algorithms, including AdaBoost, Bagging, Bernoulli Naïve Bayes (Bernoulli NB), Decision Tree (DT), Extra Tree (ET), Gaussian Naïve Bayes (Gaussian NB), Gradient Boosting, K-Nearest Neighbor (KNN), Latent

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

For peer review only

Dirichlet Allocation (LDA), Logistic Regression (LR), Multinomial Naïve Bayes (Multinomial NB), Passive Aggressive, Quadratic Discriminant Analysis (QDA), RF, Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and Ensemble Learning, were used to build models.

The model establishment was as follows. The data was standardized and divided into a training set and a test set according to 8:2. The training set was used to build models, and the test set was used to evaluate the predictive performance of the models. Ten-fold cross-validation on the training set was used for internal validation of the model, and 200 Bootstrapping samples from the test set for the evaluation of the impact of different data processing methods or machine learning algorithms on model predictive performance. Five algorithms with the largest area under curve (AUC) on each data set were used for ensemble learning.

Model evaluation

We used the AUC, accuracy, precision, recall rate, and F1 value to evaluate the predictive performance of the model. Five models with the largest AUC were compared, and the model with the best predictive performance was selected to develop an ADR prediction system of Chinese herbal injections containing Panax notoginseng saponin. SHapley Additive exPlanations (SHAP) helped to explain the contribution of variables to the model.

Sample size assessment

To evaluate the influence of different sample sizes on model predictive performance,

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

randomly extracted 10%, 20%, 30% to 100% subsets from the training set by Bootstrapping. The 10 subsets were used to establish models, respectively. Repeated the procedure 100 times and the AUC, calculated from the testing set, was used for sample size examination.

Patient and public involvement

Patients and/or the public were not directly involved in this study.

Statistical Analysis

Categorical variables were expressed as counts and percentages and continuous variables as mean ± standard deviation. Analysis of variance will be used if the data were normally distributed and the variances were equal, otherwise, Kruskal-Wallis test will be used. p value<0.05 were considered statistically significant. Hypothesis testing and Models building were implemented using the stats and sklearn packages in Python (Version3.8), respectively.

RESULTS

Research population

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.

Data cleaning

The assignment of all variables was shown in Supplementary Table 1. After data processing by 4 data filling, 5 data sampling and 3 variable selection methods, we

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

obtained 60 data sets. The results of variable selection by the Lasso and Boruta were shown in Supplementary Figure 1.

Model establishment

A total of 1080 prediction models were established by 18 machine learning algorithms and the 60 data sets. The results of ten-fold cross-validation were shown in Supplementary Table 2. Using 200 Bootstrapping samples from the test set to evaluate the impact of different data processing methods or machine learning algorithms on model predictive performance. The results showed that differences of model predictive performance exist by different data filling, data sampling, variable selection (Table 1) and machine learning algorithms (Table 2). The ensemble learning model had the best performance with an AUC of 0.793±0.083 (Table 2).

Model evaluation

The AUC, accuracy, precision, recall rate, and F1 value were used to evaluate the performance of the model. The best 5 prediction models were selected and model 1 had the best performance with an AUC of 0.9141 (Table 3). The receiver operating characteristic (ROC) curve of the 5 best model was shown in Figure 1.

Model interpretation

The importance of each variable to the final prediction model was shown in Figure 2. The result showed that pre-treatment serum levels, renal function, dermatoses, gender and age were the top 5 most important variables contributing to the model. We used the SHAP value to explain the contribution of the variables to the model, and the SHAP

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

value of the top 20 variables was shown in Figure 3. This plot explains how high and low variables values were in relation to SHAP values. According to the prediction model, the higher the SHAP value of a variable, the more likely ADR occurs.

Sample size assessment

With the continuous increased size of sample data, the AUC values of the testing sets continued to increase, which shows a sufficient sample size was included in this study (Figure 4).

Develop an ADR prediction system for Panax notoginseng saponin

According to the parameters of the best model, a prediction system for the ADR of Panax notoginseng saponin has been developed and we had obtained the software copyright. The development of ADR prediction system was shown in Figure 5. The operation and output of the system were shown in Figure 6.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Table 1 The effect of different data processing methods on model prediction performance (Bootstrapping)

		AUC		Accuracy		Precision		Recall rate		F1 value	
		Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI
Data filling											
	No filling	**0.786±0.101**	0.785-0.787	**0.770±0.070**	0.769-0.771	0.437±0.162	0.435-0.438	**0.546±0.208**	0.544-0.548	**0.460±0.142**	0.459-0.461
	Simple filling	0.687±0.094	0.686-0.688	0.761±0.076	0.760-0.761	**0.455±0.180**	0.453-0.456	0.491±0.165	0.489-0.492	0.442±0.126	0.441-0.443
	RF filling	0.677±0.095	0.676-0.678	0.759±0.077	0.758-0.760	0.446±0.181	0.444-0.447	0.488±0.162	0.487-0.490	0.440±0.129	0.439-0.441
	RF improve filling	0.678±0.092	0.677-0.678	0.756±0.077	0.755-0.757	0.443±0.179	0.442-0.445	0.485±0.161	0.483-0.486	0.435±0.125	0.434-0.436
	p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	
Data sampling											
	No sampling	**0.738±0.101**	0.737-0.739	**0.823±0.050**	0.822-0.823	**0.585±0.229**	0.583-0.588	0.390±0.178	0.388-0.391	0.441±0.172	0.439-0.442
	Random over sampler	0.718±0.109	0.717-0.719	0.765±0.070	0.764-0.765	0.437±0.154	0.435-0.438	0.531±0.189	0.529-0.533	**0.457±0.135**	0.456-0.458
	Random under sampler	0.696±0.106	0.695-0.697	0.710±0.069	0.709-0.711	0.364±0.107	0.363-0.365	**0.596±0.161**	0.594-0.597	0.441±0.109	0.440-0.442
	SMOTE over sampler	0.683±0.100	0.682-0.684	0.755±0.067	0.754-0.755	0.416±0.137	0.414-0.417	0.490±0.143	0.488-0.491	0.435±0.113	0.434-0.436
	Borderline SMOTE	0.699±0.104	0.698-0.700	0.755±0.072	0.755-0.756	0.424±0.143	0.422-0.425	0.506±0.143	0.505-0.508	0.446±0.115	0.445-0.447
	p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 20, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Variable selection										
	No selection	0.702±0.109	0.702-0.703	0.758±0.078	0.758-0.759	0.440±0.184	0.438-0.441	0.493±0.187	0.492-0.494	0.434±0.137	0.433-0.435
	Lasso selection	**0.713±0.105**	0.712-0.713	0.761±0.074	0.760-0.761	0.447±0.173	0.445-0.448	**0.513±0.177**	0.512-0.514	0.448±0.128	0.447-0.449
	Boruta selection	0.706±0.103	0.705-0.707	**0.766±0.073**	0.765-0.766	**0.449±0.170**	0.448-0.450	0.501±0.166	0.500-0.503	**0.450±0.127**	0.449-0.451
	p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	

202 AUC, Area under curve; RF, Random Forest; SMOTE, Synthetic minority oversampling technique.

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

Table 2 The effect of different machine learning algorithms on model prediction performance (Bootstrapping)

	AUC		Accuracy		Precision		Recall rate		F1 value	
	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI
machine learning algorithms										
AdaBoost	0.702±0.104	0.700-0.703	0.761±0.061	0.760-0.762	0.434±0.134	0.432-0.436	0.538±0.442	0.535-0.540	0.465±0.105	0.463-0.467
Bagging	0.749±0.083	0.748-0.750	0.776±0.064	0.774-0.777	0.457±0.137	0.454-0.459	0.486±0.159	0.483-0.489	0.452±0.112	0.450-0.454
Bernoulli NB	0.718±0.099	0.716-0.720	0.771±0.056	0.770-0.772	0.444±0.133	0.442-0.447	0.541±0.141	0.538-0.543	0.475±0.109	0.474-0.477
DT	0.667±0.085	0.665-0.668	0.738±0.067	0.737-0.739	0.388±0.127	0.386-0.390	0.491±0.151	0.489-0.494	0.417±0.105	0.416-0.419
Ensemble Learning	**0.793±0.083**	0.791-0.794	**0.810±0.058**	0.809-0.811	**0.545±0.157**	0.543-0.548	**0.576±0.162**	0.573-0.579	**0.537±0.108**	0.535-0.539
ET	0.596±0.097	0.594-0.598	0.703±0.081	0.701-0.704	0.308±0.149	0.305-0.310	0.393±0.186	0.390-0.396	0.326±0.139	0.324-0.329
Gaussian NB	0.667±0.106	0.665-0.669	0.720±0.061	0.719-0.721	0.364±0.106	0.362-0.366	0.543±0.133	0.541-0.545	0.429±0.103	0.427-0.431
Gradient Boosting	0.718±0.100	0.716-0.720	0.783±0.060	0.782-0.784	0.487±0.161	0.484-0.490	0.524±0.144	0.521-0.526	0.481±0.105	0.479-0.483
KNN	0.655±0.101	0.654-0.657	0.741±0.086	0.740-0.743	0.394±0.262	0.389-0.399	0.355±0.217	0.351-0.359	0.316±0.166	0.313-0.319
LDA	0.724±0.097	0.722-0.725	0.770±0.065	0.769-0.772	0.457±0.149	0.454-0.459	0.561±0.141	0.558-0.564	0.487±0.110	0.485-0.489
LR	0.728±0.094	0.727-0.730	0.770±0.070	0.769-0.771	0.465±0.155	0.462-0.467	0.580±0.143	0.577-0.583	0.497±0.110	0.495-0.499

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

Multinomial NB	0.727±0.099	0.725-0.728	0.753±0.071	0.752-0.754	0.450±0.170	0.447-0.453	0.570±0.175	0.567-0.573	0.467±0.111	0.465-0.469
Passive Aggressive	0.686±0.094	0.684-0.688	0.701±0.087	0.699-0.703	0.358±0.119	0.355-0.360	0.558±0.156	0.555-0.560	0.421±0.107	0.419-0.423
QDA	0.660±0.115	0.658-0.662	0.774±0.057	0.773-0.775	0.428±0.178	0.425-0.431	0.436±0.188	0.433-0.440	0.411±0.152	0.408-0.413
RF	0.742±0.088	0.741-0.744	0.792±0.075	0.791-0.793	0.534±0.194	0.531-0.538	0.430±0.175	0.427-0.432	0.444±0.119	0.441-0.446
SGD	0.720±0.099	0.718-0.722	0.762±0.064	0.761-0.764	0.452±0.196	0.448-0.455	0.507±0.213	0.503-0.511	0.434±0.141	0.432-0.437
SVM	0.735±0.090	0.734-0.737	0.792±0.073	0.790-0.793	0.533±0.194	0.529-0.536	0.443±0.165	0.440-0.446	0.449±0.115	0.447-0.451
XGBoost	0.740±0.095	0.738-0.741	0.790±0.074	0.789-0.792	0.515±0.161	0.512-0.518	0.513±0.165	0.510-0.516	0.486±0.112	0.484-0.488
p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	

Bernoulli NB, Bernoulli Naïve Bayes; DT, Decision Tree; ET, Extra Tree; Gaussian NB, Gaussian Naïve Bayes; KNN, K-Nearest Neighbor; LDA, Latent Dirichlet Allocation; LR, Logistic Regression; Multinomial NB, Multinomial Naïve Bayes; QDA, Quadratic Discriminant Analysis; SGD, Stochastic Gradient Descent; SVM, support vector machine. XGBoost, eXtreme Gradient Boosting.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

Table 3 Predictive performance indicators of the 5 best models

	AUC	accuracy	precision	recall rate	F1 value
model 1	**0.9141**	**0.8947**	**0.75**	0.6667	**0.7059**
model 2	0.9055	0.8105	0.5	**0.7778**	0.6087
model 3	0.9019	0.8421	0.6154	0.4444	0.5161
model 4	0.8997	0.8632	0.6316	0.6667	0.6486
model 5	0.8968	0.8316	0.5357	0.8333	0.6522

DISCUSSION

Traditional Chinese medicine has been used for the prevention and treatment of diseases for centuries [11]. In recent years, the application of Chinese herbal containing Panax notoginseng saponin, including injections, in clinical practice has become more and more common, while the ADR often causes concerns. Studies have shown that the Chinese herbal ingredients, traditional Chinese medicine preparation and combination medication are the important factors for the ADR of Chinese herbal injections containing Panax notoginseng saponin. Drug eruption (50.5%), allergic reactions (20.4%) and anaphylactic shock (9.7%) are the most common, and some cases are even life-threatening [5]. However, the ADR monitoring methods, including spontaneous reporting systems, prescription event monitoring and centralized hospital monitoring system, are reported after the event, and may even have data bias, underreporting or repeated reporting. Therefore, the realization of ADR prediction has important significance for prevent and avoid ADR of Chinese herbal injections containing Panax notoginseng saponin in clinical practice.

In our study, a nested case-control study was performed for data collection. Sixty

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

data sets, which were from data filling, data sampling and variable selection, were combined with 18 machine learning algorithms to establish 1080 ADR prediction models. The AUC, accuracy, precision, recall rate and F1 value were used to evaluate the predictive performance of the models. According to the parameters of the best model, an ADR prediction system for the Chinese herbal injections containing Panax notoginseng saponin was developed. This predictive system had high accuracy and precision, and had potential value for clinical application.

In recent years, some ADR prediction models based on data mining [6-9], machine learning algorithms [10, 12-15], and statistical methods [16-18], have been developed. Tangiisuran et al. [16] combined univariate analysis and multivariate binary logistic regression for the identification of clinical risk factors to develop an ADR risk model. The AUC of the model at internal and external validation stage was 0.74 and 0.73, respectively, the sensitivity was 80% and 84%, and the specificity was 55% and 43% [16]. Imai et al. [10] used artificial neural networks to predict the ADR risk and produced an AUC of 0.83. Compared with these models, the model established in our study had better predictive performance (accuracy was 0.8947, precision was 0.75, recall rate was 0.6667 and AUC was 0.914). As missing data is common in the real-world health system, the methods of data filling used in our study may be advantageous for the deal with imbalanced data in clinical real-world research. More importantly, the model with optimal predictive performance selected from the 1080 models, was used to develop the ADR risk prediction system, which is potentially convenient for clinical practice because of its' simple operation, fast calculation, and high accuracy.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

It is worth noting that Hammann et al. [19] established a decision tree model based on the chemical, physical, and structural properties of compounds for the prediction of ADR occurrence and the model had high predictive accuracies (78.9–90.2%). Unfortunately, the model ignored the effect of pathological and physiological conditions and the combination medication on ADR. More than 80 factors including the patient's pathophysiological characteristics, clinical laboratory results, and medication conditions, were performed by 3 variable selection methods in our study. Meanwhile, we using the SHAP value to explain the contribution of the variables to the model.

The importance of the variable indicates that whether the patients have dermatoses will significantly affect the models' predictive performance. Cutaneous ADR is one of the most common adverse reactions of Panax notoginseng, such as erythema multiforme, urticaria, severe erythema multiforme and acute generalized exanthematous pustulosis [20, 21]. Therefore, those patients with original dermatoses are more likely to have ADR after using Panax notoginseng. In addition, we found that the age and gender are related to the occurrence of Panax notoginseng-induced ADR, which is consistent with the results reported by Yang et al. [22].

However, our data were all from southwest China, and more data were needed to further evaluate the model prediction performance. In addition, a prospective controlled trial is required to demonstrate the accuracy of the ADR prediction system.

Contributors XWW, EWL and RST were involved in the conception and design of the study. XWW drafted the article. JYZ, HC, XWS and YLW analyzed the data.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

EWL and RST revised the manuscript. All authors gave final approval of the version to be published. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. RST is the guarantor.

Funding This study was funded by the National Natural Science Foundation of China (No. 72004020), the Key Research and Development Program of Science and Technology Department of Sichuan Province (No. 2019YFS0514), the Postgraduate Research and Teaching Reform Project of the University of Electronic Science and Technology of China (No. JYJG201919) and the Research Subject of Health Commission of Sichuan Province (No. 19PJ262).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Ethical approval: This study was approved by the Ethics Committee of Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital (2017-11-01).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Data may be obtained from a third party and are not publicly available. The first author (7190175@uestc.edu.cn) will share any publicly available data if requested by email.

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

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

For peer review only

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

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

REFERENCES

1 Xie W, Meng X, Zhai Y ,et al. Panax Notoginseng Saponins: A Review of Its Mechanisms of Antidepressant or Anxiolytic Effects and Network Analysis on Phytochemistry and Pharmacology. *Molecules* 2018; *23*.

2 Kim JH. Pharmacological and medical applications of Panax ginseng and ginsenosides: a review for use in cardiovascular diseases. *J Ginseng Res* 2018; *42*:264-269.

3 Yang F, Ma Q, Matsabisa MG ,et al. Panax notoginseng for Cerebral Ischemia: A Systematic Review. *Am J Chin Med* 2020; *48*:1331-1351.

4 Qu J, Xu N, Zhang J ,et al. Panax notoginseng saponins and their applications in nervous system disorders: a narrative review. *Ann Transl Med* 2020;

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

313 *8*:1525.

314 5 Xiang Z, Qiao T, Xiao H ,et al. The anaphylactoid constituents in Xue-Sai-
315 Tong injection. *Planta Med* 2013; *79*:1043-1050.

316 6 Wei L, Scott J. Association rule mining in the US Vaccine Adverse Event
317 Reporting System (VAERS). *Pharmacoepidemiol Drug Saf* 2015; *24*:922-933.

318 7 Harpaz R, DuMouchel W, Shah NH ,et al. Novel data-mining methodologies
319 for adverse drug event discovery and analysis. *Clin Pharmacol Ther* 2012;
320 *91*:1010-1021.

321 8 Sakaeda T, Tamon A, Kadoyama K ,et al. Data mining of the public version of
322 the FDA Adverse Event Reporting System. *Int J Med Sci* 2013; *10*:796-803.

323 9 Kadoyama K, Kuwahara A, Yamamori M ,et al. Hypersensitivity reactions to
324 anticancer agents: data mining of the public version of the FDA adverse event
325 reporting system, AERS. *J Exp Clin Cancer Res* 2011; *30*:93.

326 10 Imai S, Takekuma Y, Kashiwagi H ,et al. Validation of the usefulness of
327 artificial neural networks for risk prediction of adverse drug reactions used for
328 individual patients in clinical practice. *PLoS One* 2020; *15*:e0236789.

329 11 Liu SH, Chuang WC, Lam W ,et al. Safety surveillance of traditional Chinese
330 medicine: current and future. *Drug Saf* 2015; *38*:117-128.

331 12 Choudhury O, Park Y, Salonidis T ,et al. Predicting Adverse Drug Reactions
332 on Distributed Health Data using Federated Learning. *AMIA Annu Symp Proc*
333 2019; *2019*:313-322.

334 13 Liu X, Chen H. A research framework for pharmacovigilance in health social

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

media: Identification and evaluation of patient adverse drug event reports. *J Biomed Inform* 2015; *58*:268-279.

14 Davis J, Costa VS, Peissig P ,et al. Demand-Driven Clustering in Relational Domains for Predicting Adverse Drug Events. *Proc Int Conf Mach Learn* 2012; *2012*:1287-1294.

15 Lee CY, Chen YP. Prediction of drug adverse events using deep learning in pharmaceutical discovery. *Brief Bioinform* 2021; *22*:1884-1901.

16 Tangiisuran B, Scutt G, Stevenson J ,et al. Development and validation of a risk model for predicting adverse drug reactions in older people during hospital stay: Brighton Adverse Drug Reactions Risk (BADRI) model. *PLoS One* 2014; *9*:e111254.

17 Clothier HJ, Lawrie J, Lewis G ,et al. SAEFVIC: Surveillance of adverse events following immunisation (AEFI) in Victoria, Australia, 2018. *Commun Dis Intell (2018)* 2020; *44*.

18 Alvarez Y, Hidalgo A, Maignen F ,et al. Validation of statistical signal detection procedures in eudravigilance post-authorization data: a retrospective evaluation of the potential for earlier signalling. *Drug Saf* 2010; *33*:475-487.

19 Hammann F, Gutmann H, Vogt N ,et al. Prediction of adverse drug reactions using decision tree modeling. *Clin Pharmacol Ther* 2010; *88*:52-59.

20 Yan S, Xiong H, Shao F ,et al. HLA-C*12:02 is strongly associated with Xuesaitong-induced cutaneous adverse drug reactions. *Pharmacogenomics J* 2019; *19*:277-285.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

21 Chen WJ, Kuang YY, Li JT. Analysis on 13 Cases of Adverse Drug Reaction by Xuesaitong Injection. *Journal of North Pharmacy* 2013; *10*:16-17.

22 Yang P, Qian N, Yao D ,et al. 62 Cases of Adverse Reactions in Xuesaitong Oral Preparations. *Chinese Medicine Modern Distance Education of China* 2021; *19*:34-36.

Figure 1 ROC curve of the 5 best models.

Figure 2 Importance matrix plot of each variable to the final prediction model. Variable names were shown in Supplementary Table 1. X83, pre-treatment serum levels; X55, renal function; X25, dermatoses; X1, gender; X2, age; X29, dose; X62, low-density lipoprotein; X64, hypoproteinemia; X30, anti-infective agents; X82, pre-treatment indicators of carcinoma; X79, hemoglobin; X6, history of allergy; X16, respiratory diseases; X66, albumin/globulin; X78, red blood cell; X81, hypersensitive C-reactive protein; X51, dermatology medication; X77, eosinophils; X13, Charlson comorbidity index (Score); X57, serum potassium.

Figure 3 SHAP summary plot of the top 20 variables of the model. Red represents higher variable values, and blue represents lower variable values. Variable names were shown in Supplementary Table 1. X83, pre-treatment serum levels; X55, renal function; X25, dermatoses; X1, gender; X2, age; X29, dose; X62, low-density lipoprotein; X64, hypoproteinemia; X30, anti-infective agents; X82, pre-treatment indicators of carcinoma; X79, hemoglobin; X6, history of allergy; X16, respiratory diseases; X66, albumin/globulin; X78, red blood cell; X81, hypersensitive C-reactive

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

protein; X51, dermatology medication; X77, eosinophils; X13, Charlson comorbidity index (Score); X57, serum potassium.

Figure 4 Sample size validation. The vertical bars represent the 95% confidence interval (CI) of AUC of ROC.

Figure 5 The development of ADR prediction system.

Figure 6 The operation (A) and output (B) of the ADR prediction system.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

385

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

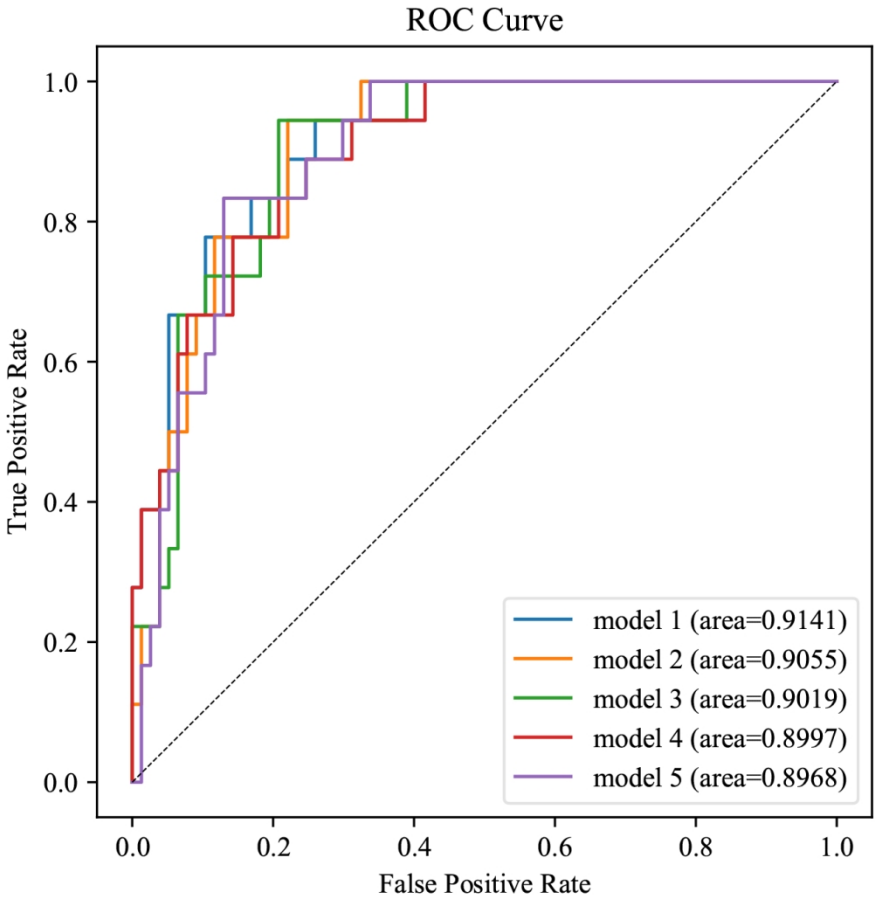

Figure 1 ROC curve of the 5 best models.

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

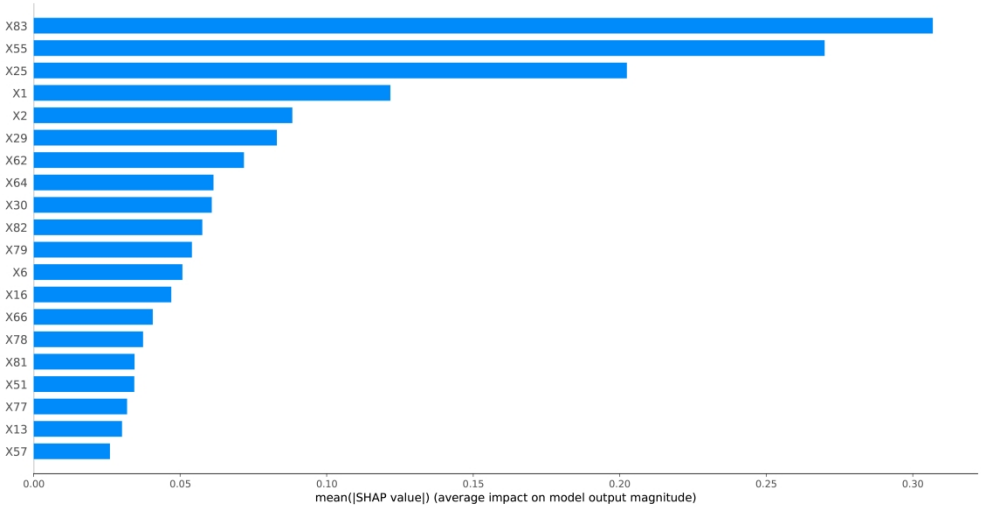

Figure 2 Importance matrix plot of each variable to the final prediction model.

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

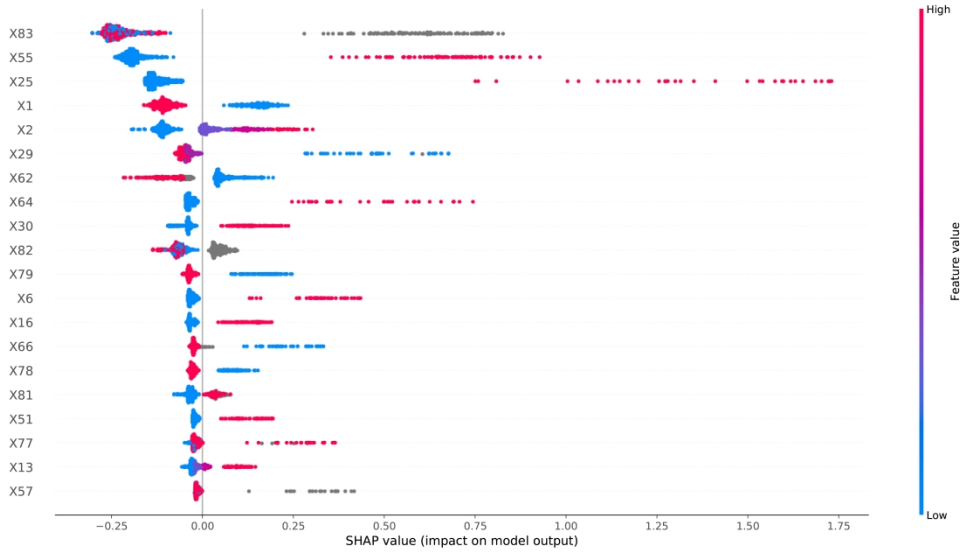

Figure 3 SHAP summary plot of the top 20 variables of the model.

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

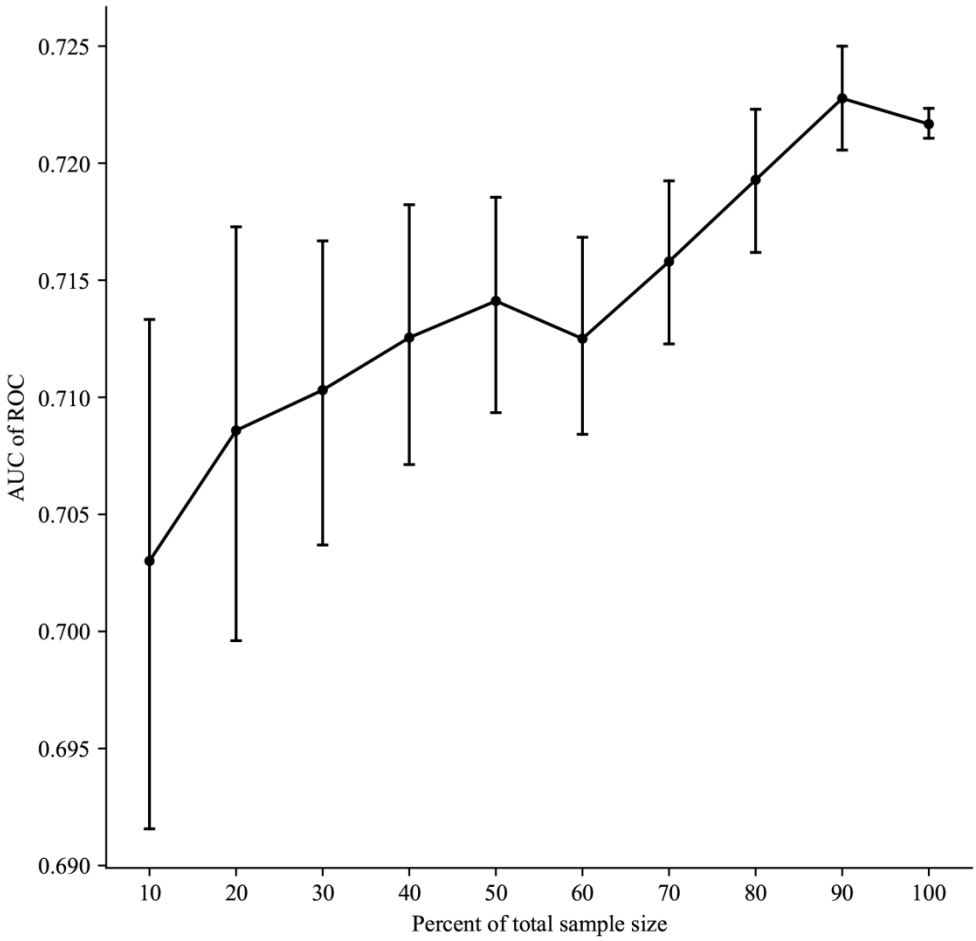

Figure 4 Sample size validation.

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

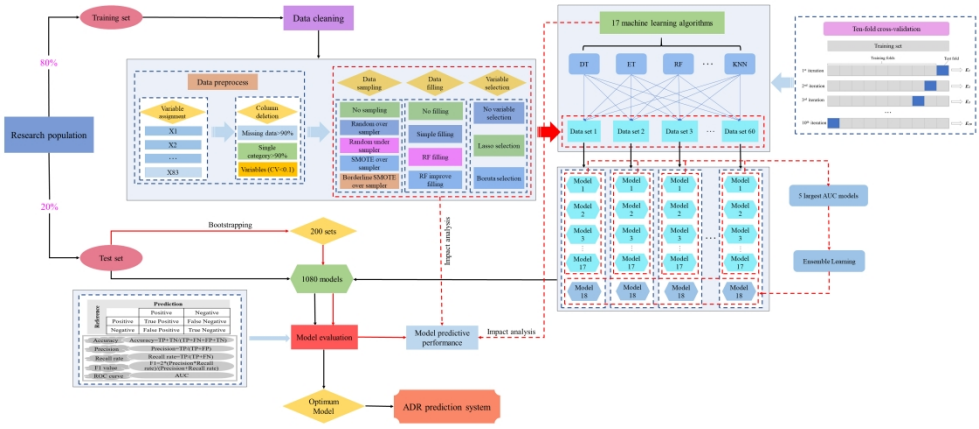

Figure 5 The development of ADR prediction system.

1050x472mm (96 x 96 DPI)

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

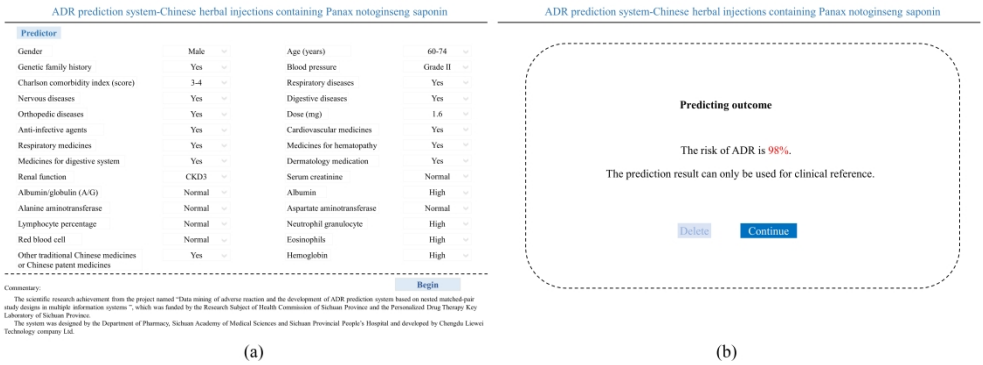

Figure 6 The operation (A) and output (B) of the ADR prediction system.

1899x688mm (96 x 96 DPI)

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

Table 1 Variable assignment

Number	Variable	Assignment
	Adverse drug reaction	1, Yes; 0, No
X1	Gender	1, Male; 0, Female
X2	Age (years)	1, ≤ 44; 2, 45 ≤ Age ≤ 59; 3, 60 ≤ Age ≤ 74; 4, ≥ 75
X3	Body mass index (BMI, kg/m^2)	1, < 18.5; 2, 18.5 ≤ BMI ≤ 23.9; 3, ≥ 24
X4	Asians	1, Yes; 0, No
X5	Genetic family history	1, Yes; 0, No
X6	History of allergy	1, Yes; 0, No
X7	Smoking	1, Yes; 0, No
X8	Alcohol	1, Yes; 0, No
X9	Temperature (℃)	1, < 36.1; 2, 36.1 ≤ Temperature ≤ 37.2; 3, > 37.3
X10	Pulse (beats/min)	1, < 60; 2, 60 ≤ Pulse ≤ 100, 3, > 100
X11	Breathe (times/min)	1, < 12; 2, 12 ≤ Breathe ≤ 20; 3, > 20
X12	Blood pressure	0, Normal (systolic pressure ≤ 139 mmHg or diastolic pressure ≤ 89 mmHg); 1, Grade I (140 mmHg ≤ systolic pressure ≤ 159 mmHg or 90 mmHg ≤ diastolic pressure ≤ 99 mmHg); 2, Grade II (160 mmHg ≤ systolic pressure ≤ 179 mmHg or 100 mmHg ≤ diastolic pressure ≤ 109 mmHg); 3, Grade III (systolic pressure ≥180 mmHg or diastolic pressure ≥110 mmHg)
X13	Charlson comorbidity index (Score)	1, 0; 2, 1 or 2; 3, 3 or 4; 4, ≥ 5
X14	Cardiovascular disease	1, Yes; 0, No

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

X15	Endocrine diseases	1, Yes; 0, No
X16	Respiratory diseases	1, Yes; 0, No
X17	Nervous diseases	1, Yes; 0, No
X18	Digestive diseases	1, Yes; 0, No
X19	Neoplastic diseases	1, Yes; 0, No
X20	Orthopedic diseases	1, Yes; 0, No
X21	Genito-urinary diseases	1, Yes; 0, No
X22	Hematopathy	1, Yes; 0, No
X23	Oculopathy	1, Yes; 0, No
X24	Ear-nose-throat diseases	1, Yes; 0, No
X25	Dermatoses	1, Yes; 0, No
X26	Immune rheumatism	1, Yes; 0, No
X27	Other diseases	1, Yes; 0, No
X28	Solvent	1, 0.9% sodium chloride injection; 2, 5% glucose injection; 3, Other solvents
X29	Dose (mg)	1, < 1.6; 2, =1.6; 3, > 1.6
X30	Anti-infective agents	1, Yes; 0, No
X31	Cardiovascular medicines	1, Yes; 0, No
X32	Medicines for digestive system	1, Yes; 0, No
X33	Respiratory medicines	1, Yes; 0, No
X34	Nervous system medicines	1, Yes; 0, No
X35	Medication in mental disorders	1, Yes; 0, No

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

X36	Non-steroidal anti-inflammatory drugs	1, Yes; 0, No
X37	Antiallergic agent	1, Yes; 0, No
X38	Genito-urinary system medicines	1, Yes; 0, No
X39	Medicines for hematopathy	1, Yes; 0, No
X40	Endocrine agents or hormone drugs	1, Yes; 0, No
X41	Antineoplastic drugs	1, Yes; 0, No
X42	Amino acids, vitamins, minerals or other nutrition preparations	1, Yes; 0, No
X43	Regulating water, electrolyte or acid-base balance drugs	1, Yes; 0, No
X44	Adjuvant agents to anesthesia or anesthetics	1, Yes; 0, No
X45	Diagnostic agents	1, Yes; 0, No
X46	Biological agents	1, Yes; 0, No
X47	Obstetrical-gynecological drugs	1, Yes; 0, No
X48	Stomatological preparations	1, Yes; 0, No
X49	Ophthalmic medication	1, Yes; 0, No
X50	Ear-nose-throat medication	1, Yes; 0, No
X51	Dermatology medication	1, Yes; 0, No
X52	Other traditional Chinese medicines	1, Yes; 0, No

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	or Chinese patent medicines	
X53	Urea	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X54	Serum creatinine	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X55	Renal function	1, Glomerular filtration rate ≥ 90 ml/(min·1.73m^2); 2, 60ml/(min·1.73m^2) ≤ Glomerular filtration rate ≤ 89ml/(min·1.73m^2); 3, 30ml/(min·1.73m^2) ≤ Glomerular filtration rate ≤59 ml/(min·1.73m^2); 4, 15ml/(min·1.73m^2) ≤ Glomerular filtration rate ≤29 ml/(min·1.73m^2); 5, Glomerular filtration rate < 15 ml/(min·1.73m^2)
X56	Blood glucose	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X57	Serum potassium	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X58	Serum sodium	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X59	Total cholesterol	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X60	Triglyceride	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X61	High-density lipoprotein	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X62	Low-density lipoprotein	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X63	Albumin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X64	Hypoproteinemia	1, Yes; 0, No
X65	Globulin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X66	Albumin/globulin (A/G)	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X67	Aspartate aminotransferase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X68	Alanine aminotransferase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

X69	Liver function	1, Less than 3 times upper limit of normal range of liver function tests (ULN of LFTs); 2, 3~5 times ULN of LFTs; 3, More than 5 times ULN of LFTs
X70	Total bilirubin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X71	Lactic dehydrogenase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X72	Creatine kinase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X73	White blood cell	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X74	Neutrophil granulocyte	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X75	Lymphocyte percentage	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X76	Monocyte percentage	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X77	Eosinophils	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X78	Red blood cell	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X79	Hemoglobin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X80	Platelet count	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X81	Hypersensitive C-reactive protein	0, Within the normal range; 1, Above the normal range
X82	Pre-treatment indicators of carcinoma	0, Within the normal range; 1, Above the normal range
X83	Pre-treatment serum levels	0, Within the normal range; 1, Above the normal range

bmjopen-2022-061457 on 18 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Table 2 The effect of different data processing methods and machine learning algorithms on model prediction performance (Ten-fold cross-validation)

		AUC		Accuracy		Precision		Recall rate		F1 value	
		Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI
Data filling											
	No filling	0.868±0.099	0.864-0.872	0.820±0.093	0.816-0.823	0.772±0.190	0.765-0.779	0.720±0.254	0.710-0.730	0.729±0.217	0.721-0.737
	Simple filling	0.881±0.097	0.877-0.885	0.828±0.100	0.824-0.832	0.793±0.165	0.787-0.799	0.746±0.243	0.737-0.756	0.751±0.197	0.744-0.759
	RF filling	0.885±0.095	0.881-0.888	0.831±0.095	0.827-0.835	**0.802±0.157**	0.796-0.808	0.749±0.237	0.740-0.759	**0.757±0.189**	0.750-0.764
	RF improve filling	**0.887±0.094**	0.883-0.890	**0.832±0.096**	0.828-0.835	0.799±0.158	0.793-0.806	**0.751±0.240**	0.742-0.760	0.757±0.191	0.749-0.764
	p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	
Data sampling											
	No sampling	0.824±0.088	0.820-0.828	0.832±0.050	0.830-0.835	0.641±0.271	0.629-0.653	0.399±0.197	0.391-0.408	0.464±0.193	0.455-0.472
	Random over sampler	**0.923±0.063**	0.920-0.925	0.858±0.085	0.854-0.861	**0.849±0.079**	0.845-0.852	0.872±0.118	0.867-0.877	0.857±0.089	0.854-0.861
	Random under sampler	0.815±0.107	0.810-0.819	0.732±0.104	0.728-0.737	0.783±0.145	0.776-0.789	0.678±0.188	0.670-0.686	0.707±0.132	0.701-0.713
	SMOTE over sampler	0.920±0.072	0.917-0.923	0.857±0.081	0.853-0.860	0.844±0.071	0.841-0.848	0.875±0.125	0.869-0.880	0.856±0.089	0.852-0.860
	Borderline SMOTE	0.919±0.077	0.916-0.923	**0.859±0.085**	0.855-0.862	0.841±0.074	0.837-0.844	**0.885±0.130**	0.879-0.890	**0.859±0.093**	0.855-0.863
	p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	
Variable selection											

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	No selection	0.870±0.105	0.867-0.874	0.820±0.104	0.817-0.824	0.780±0.178	0.774-0.786	0.733±0.254	0.725-0.742	0.737±0.208	0.730-0.744
	Lasso selection	**0.889±0.089**	0.886-0.892	**0.835±0.090**	0.832-0.838	**0.801±0.165**	0.796-0.807	**0.751±0.240**	0.743-0.759	**0.758±0.196**	0.752-0.765
	Boruta selection	0.881±0.094	0.878-0.884	0.827±0.093	0.824-0.830	0.794±0.162	0.788-0.799	0.741±0.236	0.733-0.749	0.750±0.191	0.744-0.757
	p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	
machine learning algorithms											
	AdaBoost	0.871±0.092	0.864-0.879	0.813±0.093	0.806-0.820	0.784±0.136	0.773-0.795	0.731±0.202	0.715-0.747	0.745±0.160	0.733-0.758
	Bagging	0.907±0.102	0.898-0.915	0.854±0.101	0.846-0.863	0.805±0.158	0.793-0.818	0.791±0.245	0.771-0.810	0.785±0.196	0.769-0.801
	Bernoulli NB	0.866±0.082	0.860-0.873	0.802±0.085	0.795-0.809	0.771±0.144	0.759-0.783	0.719±0.178	0.705-0.733	0.736±0.148	0.724-0.748
	DT	0.815±0.110	0.806-0.824	0.805±0.089	0.797-0.812	0.773±0.158	0.760-0.786	0.715±0.237	0.696-0.734	0.724±0.184	0.709-0.739
	ET	0.829±0.110	0.821-0.838	0.809±0.092	0.801-0.816	0.767±0.164	0.754-0.780	0.714±0.255	0.694-0.735	0.720±0.207	0.704-0.737
	Gaussian NB	0.845±0.089	0.838-0.852	0.786±0.085	0.779-0.793	0.734±0.155	0.722-0.747	0.743±0.164	0.730-0.756	0.730±0.143	0.719-0.742
	Gradient Boosting	0.891±0.102	0.883-0.899	0.841±0.099	0.833-0.849	0.822±0.149	0.810-0.834	0.746±0.252	0.725-0.766	0.762±0.194	0.747-0.778
	KNN	0.896±0.084	0.890-0.903	0.830±0.098	0.822-0.838	0.747±0.296	0.724-0.771	0.687±0.381	0.656-0.717	0.674±0.326	0.648-0.700
	LDA	0.897±0.073	0.891-0.903	0.835±0.081	0.829-0.842	0.805±0.117	0.796-0.815	0.768±0.191	0.753-0.783	0.777±0.144	0.765-0.788
	LR	0.893±0.076	0.886-0.899	0.834±0.082	0.827-0.840	0.815±0.119	0.805-0.824	0.754±0.216	0.737-0.772	0.767±0.157	0.755-0.780
	Multinomial NB	0.839±0.071	0.834-0.845	0.773±0.078	0.766-0.779	0.753±0.161	0.740-0.766	0.653±0.235	0.634-0.672	0.676±0.190	0.660-0.691
	Passive Aggressive	0.836±0.098	0.828-0.844	0.780±0.091	0.772-0.787	0.723±0.161	0.711-0.736	0.720±0.205	0.703-0.736	0.712±0.172	0.698-0.725

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

QDA	0.915±0.081	0.909-0.922	0.860±0.089	0.853-0.868	0.827±0.152	0.814-0.839	0.798±0.184	0.783-0.812	0.805±0.156	0.792-0.817
RF	0.919±0.097	0.911-0.926	0.871±0.100	0.863-0.879	0.843±0.154	0.831-0.856	0.775±0.268	0.753-0.796	0.788±0.214	0.771-0.805
SGD	0.895±0.075	0.889-0.901	0.832±0.082	0.825-0.839	0.803±0.197	0.787-0.819	0.710±0.287	0.687-0.733	0.726±0.238	0.707-0.745
SVM	**0.926±0.086**	0.919-0.933	**0.875±0.096**	0.867-0.883	**0.858±0.144**	0.847-0.870	0.776±0.271	0.754-0.797	0.791±0.217	0.773-0.808
XGBoost	0.922±0.092	0.914-0.929	0.869±0.100	0.861-0.877	0.825±0.153	0.812-0.837	**0.810±0.229**	0.792-0.828	**0.808±0.185**	0.793-0.822
p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	

AUC, Area under curve; RF, Random Forest; SMOTE, Synthetic minority oversampling technique; Bernoulli NB, Bernoulli Naïve Bayes; DT, Decision Tree; ET, Extra Tree; Gaussian NB, Gaussian Naïve Bayes; KNN, K-Nearest Neighbor; LDA, Latent Dirichlet Allocation; LR, Logistic Regression; Multinomial NB, Multinomial Naïve Bayes; QDA, Quadratic Discriminant Analysis; SGD, Stochastic Gradient Descent; SVM, support vector machine. XGBoost, eXtreme Gradient Boosting.

For Peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

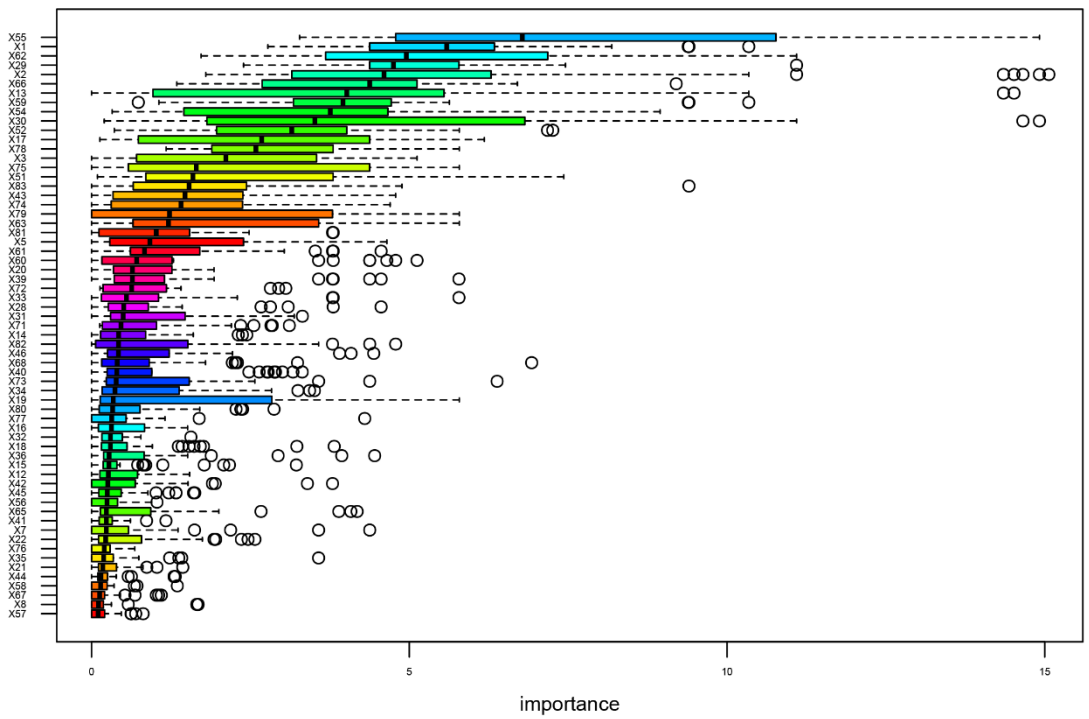

Figure 1 Variable selection by Lasso and Boruta. Variable names were shown in Table S1.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Reporting checklist for prediction model development/validation.

Based on the TRIPOD guidelines.

Instructions to authors

Complete this checklist by entering the page numbers from your manuscript where readers will find each of the items listed below.

Your article may not currently address all the items on the checklist. Please modify your text to include the missing information. If you are certain that an item does not apply, please write "n/a" and provide a short explanation.

Upload your completed checklist as an extra file when you submit to a journal.

In your methods section, say that you used the TRIPODreporting guidelines, and cite them as:

Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement.

		Reporting Item	Page Number
Title			
	#1	Identify the study as developing and / or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract			

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only

	#2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2
Introduction			
	#3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	3
	#3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	4
Methods			
Source of data	#4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	5
Source of data	#4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	5
Participants	#5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5
Participants	#5b	Describe eligibility criteria for participants.	5
Participants	#5c	Give details of treatments received, if relevant	5

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

Outcome	#6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	7
Outcome	#6b	Report any actions to blind assessment of the outcome to be predicted.	7
Predictors	#7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured	6
Predictors	#7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	6
Sample size	#8	Explain how the study size was arrived at.	5
Missing data	#9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	6
Statistical analysis methods	#10a	If you are developing a prediction model describe how predictors were handled in the analyses.	6
Statistical analysis methods	#10b	If you are developing a prediction model, specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	7
Statistical analysis methods	#10c	If you are validating a prediction model, describe how the predictions were calculated.	7
Statistical analysis methods	#10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	7

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only

Statistical analysis methods	#10e	If you are validating a prediction model, describe any model updating (e.g., recalibration) arising from the validation, if done	7
Risk groups	#11	Provide details on how risk groups were created, if done.	7
Development vs. validation	#12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	7
Results			
Participants	#13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	8
Participants	#13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	8
Participants	#13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	8
Model development	#14a	If developing a model, specify the number of participants and outcome events in each analysis.	9
Model development	#14b	If developing a model, report the unadjusted association, if calculated between each candidate predictor and outcome.	9

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

Model specification	#15a	If developing a model, present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	9
Model specification	#15b	If developing a prediction model, explain how to the use it.	9
Model performance	#16	Report performance measures (with CIs) for the prediction model.	9
Model-updating	#17	If validating a model, report the results from any model updating, if done (i.e., model specification, model performance).	9
Discussion			
Limitations	#18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	17
Interpretation	#19a	For validation, discuss the results with reference to performance in the development data, and any other validation data	15
Interpretation	#19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	15
Implications	#20	Discuss the potential clinical use of the model and implications for future research	16

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only

Other information

Supplementary information	#21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	18
Funding	#22	Give the source of funding and the role of the funders for the present study.	18

None The TRIPOD checklist is distributed under the terms of the Creative Commons Attribution License CC-BY. This checklist can be completed online using https://www.goodreports.org/, a tool made by the EQUATOR Network in collaboration with Penelope.ai

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

BMJ Open

Develop an ADR prediction system of Chinese herbal injections containing Panax notoginseng saponin: a nested case-control study using machine learning

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2022-061457.R1
Article Type:	Original research
Date Submitted by the Author:	19-Jul-2022
Complete List of Authors:	Wu, Xing-Wei; University of Electronic Science and Technology of China Sichuan Provincial People's Hospital, Pharmacy; Chinese Academy of Sciences Sichuan Translational Medicine Research Hospital Zhang, Jia-Ying ; Chengdu First People's Hospital, Pharmacy Chang, Huan ; University of Electronic Science and Technology of China Sichuan Provincial People's Hospital, Pharmacy Song, Xue-Wu ; University of Electronic Science and Technology of China Sichuan Provincial People's Hospital, Pharmacy; Chinese Academy of Sciences Sichuan Translational Medicine Research Hospital Wen, Ya-Lin ; University of Electronic Science and Technology of China Sichuan Provincial People's Hospital, Pharmacy Long, En-Wu ; University of Electronic Science and Technology of China Sichuan Provincial People's Hospital, Pharmacy; Chinese Academy of Sciences Sichuan Translational Medicine Research Hospital Tong, Rong-Sheng; University of Electronic Science and Technology of China Sichuan Provincial People's Hospital, Pharmacy; Chinese Academy of Sciences Sichuan Translational Medicine Research Hospital
Primary Subject Heading:	Medical management
Secondary Subject Heading:	Medical management
Keywords:	Adverse events < THERAPEUTICS, Herbal medicine < THERAPEUTICS, Toxicity < THERAPEUTICS

SCHOLARONE™
Manuscripts

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

For peer review only

Develop an ADR prediction system of Chinese herbal injections containing Panax notoginseng saponin: a nested case-control study using machine learning

Xing-Wei Wu[1,2], Jia-Ying Zhang[3], Huan Chang[1], Xue-Wu Song[1,2], Ya-Lin Wen[1], En-Wu Long[1,2], Rong-Sheng Tong[1,2]

[1]Department of Pharmacy, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, China,

[2]Chinese Academy of Sciences Sichuan Translational Medicine Research Hospital, Chengdu 610072, China,

[3]Department of Pharmacy, Chengdu First People's Hospital, Chengdu 610095, China

Correspondence to

Dr Rong-Sheng Tong, Department of Pharmacy, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, China. Chinese Academy of Sciences Sichuan Translational Medicine Research Hospital, Chengdu 610072, China. E-mail: 318004031@qq.com

Word count: 2349

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Develop an ADR prediction system of Chinese herbal injection containing Panax notoginseng saponin: a nested case-control study using machine learning

ABSTRACT

Objective This study aimed 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.

Design A nested case-control study.

Setting National Center for ADR Monitoring and the Electronic Medical Record (EMR) system.

Participants All patients were from 5 medical institutions in Sichuan Province from January 2010 to December 2018.

Main outcomes/measures Data of patients with ADR who used Chinese herbal injections containing Panax notoginseng saponin was collected from the National Center for ADR Monitoring. A nested case-control study was used to randomly match patients without ADR from the EMR system by the ratio of 1:4. Eighteen machine learning algorithms were applied for the development of ADR prediction models. Area under curve (AUC), accuracy, precision, recall rate and F1 value were used to evaluate the predictive performance of the model. An ADR prediction system was established by the best model selected from the 1080 models.

Results A total of 530 patients from 5 medical institutions were included, and 1080

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

ADR prediction models were developed. Among these models, the AUC of the best capable one was 0.9141 and the accuracy was 0.8947. According to the best model, a prediction system, which can provide early identification of patients at risk for the ADR of Panax notoginseng saponin, has been established.

Conclusion The prediction system developed based on the machine learning model in this study had good predictive performance and potential clinical application.

Key words Adverse drug reactions, Chinese herbal injection, Machine learning, Prediction system, Panax notoginseng saponin

Strengths and limitations of this study

- To the best of our knowledge, this study was the first to develop an ADR prediction system for Chinese herbal injection containing Panax notoginseng saponin using machine learning.
- Data of ADR patients came from the National Center for Adverse Drug Reaction Monitoring, which is highly representative.
- In order to obtain the best model, the data processing adopted 4 data filling, 5 data sampling, 3 variable selection methods, and 18 machine learning algorithms were applied for model establishment.
- The area under curve, accuracy, precision, recall rate, and F1 value were used to evaluate the predictive performance of the model.
- As the study population was all from southwest China, the results may be biased while the prediction system was applied in other medical institutions.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

INTRODUCTION

Panax notoginseng saponins, as the main ingredients of Panax notoginseng (Buck.) F.H.Chen, has been widely used in the disease therapy of nervous system and cardio-cerebral vascular system [1-4]. High frequency of adverse drug reactions (ADR) in Chinese herbal containing Panax notoginseng saponin has received widespread attention. Among these ADR, about 69.57% were caused by injections, mainly manifested as drug eruption (50.5%), allergic reaction (20.4%) and anaphylactic shock (9.7%), which can be life-threatening in severe cases [5].

At present, ADR is mainly monitored by spontaneous reporting system, case-control study, cohort study, prescription event monitoring and centralized hospital monitoring system. However, most of these methods have obvious hysteresis. Therefore, there is an increasing need to develop an ADR antecedent prediction system to prevent ~~and avoid~~ the occurrence of ADR in Chinese herbal injections containing Panax notoginseng saponin.

Machine learning, the core technology of artificial intelligence, is commonly used to build prediction models. In recent years, some prediction models for ADR have been established [6-10]. Based on a clustering method for the postprocessing of association rules, Lai et al. [6] developed an application of stepwise association rule mining to identify the associations between vaccine and multiple adverse events. In addition, Imai et al. [10] used artificial neural networks to evaluate vancomycin-induced nephrotoxicity. However, small sample size, incomplete patient information, and unsatisfactory predictive performance restrict the application of ADR prediction models in clinical

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

practice. In view of these challenges, this study aimed to develop an ADR prediction system of Chinese herbal injections containing Panax notoginseng saponin based on machine learning algorithms and provide reference for clinical ADR management and prevention.

METHODS

Data collection

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 Electronic Medical Record (EMR) system of the 5 medical institutions. The ratio of patients with ADR to those without ADR was 1:4. For multiple lab results, in order to facilitate clinical application, we selected the last results of patients before the usage of medication. And for multiple admissions, all patients were included according to their first admission.

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.

Data cleaning

Variable assignment

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Binary-state variables were directly assigned values of 0 or 1. According to whether in the normal range, clinical laboratory variables were assigned values of 1, 2 and 3 (1, below the normal range; 2, within the normal range; and 3, above the normal range).

Column deletion

Variables with missing data >90%, or a single category >90%, or the coefficient of variation (CV) <0.1 were deleted.

Data filling

There are 4 ways to data filling. No filling: retained the original data. Simple filling: missing data of continuous variables replaced by the mean or median, and categorical variables by the mode. Random Forest (RF) filling: used the RF model to predict and replace the missing data directly. RF improve filling: ordered variables based on the number of missing data that were replaced by RF filling next.

Data sampling

No sampling: built models from the original data. Random over sampler: randomly replicated the data of fewer categories to match the sample size to that of more categories. Random under sampler: deleted the data of more categories to match the sample size to that of fewer categories. Synthetic minority oversampling technique (SMOTE) over sampler: synthesize new data from a small amount of original data. Borderline SMOTE over sampler: synthesize new data from borderline data.

Variable selection

No variable selection or use Lasso or Boruta for variable selection.

Model establishment

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Through different data filling, data sampling and variable selection, 60 data sets were obtained. Eighteen machine learning algorithms, including AdaBoost, Bagging, Bernoulli Naïve Bayes (Bernoulli NB), Decision Tree (DT), Extra Tree (ET), Gaussian Naïve Bayes (Gaussian NB), Gradient Boosting, K-Nearest Neighbor (KNN), Latent Dirichlet Allocation (LDA), Logistic Regression (LR), Multinomial Naïve Bayes (Multinomial NB), Passive Aggressive, Quadratic Discriminant Analysis (QDA), RF, Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and Ensemble Learning, were used to build models.

The model establishment was as follows. The data were randomly divided 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. Ten-fold cross-validation on the training set was applied for internal validation of the model, and 200 Bootstrapping samples from the test set for the evaluation of the impact of different data processing methods or machine learning algorithms on model predictive performance. Ensemble learning models were developed by 5 machine learning algorithms with the largest area under curve (AUC) on each data set.

Model evaluation

We used the AUC, accuracy, precision, recall rate, and F1 value to evaluate the predictive performance of the model. Five models with the largest AUC were compared, and the best model was selected to develop an ADR prediction system of Chinese herbal injections containing Panax notoginseng saponin. SHapley Additive exPlanations (SHAP) helped to explain the contribution of variables to the model.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Sample size assessment

To evaluate the influence of different sample sizes on model predictive performance, randomly extracted 10%, 20%, 30% to 100% subsets from the training set by Bootstrapping. The 10 subsets were used to establish models, respectively. Repeated the procedure 100 times and the AUC, calculated from the testing set, was used for sample size examination.

Patient and public involvement

Patients and/or the public were not directly involved in this study.

Statistical Analysis

Categorical variables were expressed as counts and percentages and continuous variables as mean ± standard deviation. Analysis of variance will be used if the data were normally distributed and the variances were equal, otherwise, Kruskal-Wallis test will be used. p value<0.05 were considered statistically significant. Hypothesis testing and models building were implemented using the stats and sklearn packages in Python (Version3.8), respectively.

RESULTS

Research population

A total of 530 patients were enrolled in this study, of which 106 patients had ADR. The patients included 250 (47.17%) males and 280 (52.83%) females. The demographic and clinical characteristics of the patients were shown in Supplementary Table 1.

Data cleaning

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

The results of 83 variables assignment were shown in Supplementary Table 2. After the column deletion, 63 variables were included in the following study (Supplementary Table 3). Then, 4 data filling methods were used for replacing the 1,290 (3.86%) missing data. We used Lasso or Boruta for variable selection, and the results were shown in Supplementary Table 3. Using 4 data filling, 5 data sampling and 3 variable selection methods for data processing respectively, 60 data sets were obtained.

Model establishment

A total of 1080 prediction models were established by 18 machine learning algorithms and 60 data sets. The results of ten-fold cross-validation were shown in Supplementary Table 4. Using 200 Bootstrapping samples from the test set to evaluate the impact of different data processing methods or machine learning algorithms on model predictive performance. The results showed that differences of model predictive performance exist by different data filling, data sampling, variable selection (Table 1) and machine learning algorithms (Table 2). The ensemble learning model had the best performance with an AUC of 0.793±0.083 (Table 2).

Model evaluation

The AUC, accuracy, precision, recall rate, and F1 value were used to evaluate the performance of the model. The best 5 models were selected and model 1 had the best performance with an AUC of 0.9141 (Table 3). The receiver operating characteristic (ROC) curve of the 5 best models were shown in Figure 1.

Model interpretation

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

The importance of each variable to the final prediction model was shown in Figure 2. The result showed that pre-treatment serum levels, renal function, dermatoses, gender and age were the top 5 most important variables for the model. We used the SHAP value to explain the contribution of the variables to the model, and the SHAP value of the top 20 was shown in Figure 3. This plot explains how high and low variables values were in relation to SHAP values. For the prediction model, the higher the SHAP value of a variable, the more likely ADR occurs.

Sample size assessment

With the continuously increased size of sample data, the AUC values of the testing sets continued to increase, which shows a sufficient sample size included in this study (Figure 4).

Develop an ADR prediction system for Panax notoginseng saponin

According to the best model, a prediction system for the ADR of Panax notoginseng saponin has been developed and we had obtained the software copyright. The development of the ADR prediction system was shown in Figure 5. The operation and output of the system were shown in Figure 6.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 20, 2024 by guest. Protected by copyright.

205 **Table 1** The effect of different data processing methods on model prediction performance (Bootstrapping)

		AUC		Accuracy		Precision		Recall rate		F1 value	
		Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI
Data filling											
	No filling	**0.786±0.101**	0.785-0.787	**0.770±0.070**	0.769-0.771	0.437±0.162	0.435-0.438	**0.546±0.208**	0.544-0.548	**0.460±0.142**	0.459-0.461
	Simple filling	0.687±0.094	0.686-0.688	0.761±0.076	0.760-0.761	**0.455±0.180**	0.453-0.456	0.491±0.165	0.489-0.492	0.442±0.126	0.441-0.443
	RF filling	0.677±0.095	0.676-0.678	0.759±0.077	0.758-0.760	0.446±0.181	0.444-0.447	0.488±0.162	0.487-0.490	0.440±0.129	0.439-0.441
	RF improve filling	0.678±0.092	0.677-0.678	0.756±0.077	0.755-0.757	0.443±0.179	0.442-0.445	0.485±0.161	0.483-0.486	0.435±0.125	0.434-0.436
	p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	
Data sampling											
	No sampling	**0.738±0.101**	0.737-0.739	**0.823±0.050**	0.822-0.823	**0.585±0.229**	0.583-0.588	0.390±0.178	0.388-0.391	0.441±0.172	0.439-0.442
	Random over sampler	0.718±0.109	0.717-0.719	0.765±0.070	0.764-0.765	0.437±0.154	0.435-0.438	0.531±0.189	0.529-0.533	**0.457±0.135**	0.456-0.458
	Random under sampler	0.696±0.106	0.695-0.697	0.710±0.069	0.709-0.711	0.364±0.107	0.363-0.365	**0.596±0.161**	0.594-0.597	0.441±0.109	0.440-0.442
	SMOTE over sampler	0.683±0.100	0.682-0.684	0.755±0.067	0.754-0.755	0.416±0.137	0.414-0.417	0.490±0.143	0.488-0.491	0.435±0.113	0.434-0.436
	Borderline SMOTE	0.699±0.104	0.698-0.700	0.755±0.072	0.755-0.756	0.424±0.143	0.422-0.425	0.506±0.143	0.505-0.508	0.446±0.115	0.445-0.447
	p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Variable selection										
	No selection	0.702±0.109	0.702-0.703	0.758±0.078	0.758-0.759	0.440±0.184	0.438-0.441	0.493±0.187	0.492-0.494	0.434±0.137	0.433-0.435
	Lasso selection	**0.713±0.105**	0.712-0.713	0.761±0.074	0.760-0.761	0.447±0.173	0.445-0.448	**0.513±0.177**	0.512-0.514	0.448±0.128	0.447-0.449
	Boruta selection	0.706±0.103	0.705-0.707	**0.766±0.073**	0.765-0.766	**0.449±0.170**	0.448-0.450	0.501±0.166	0.500-0.503	**0.450±0.127**	0.449-0.451
	p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	

206 AUC, Area under curve; RF, Random Forest; SMOTE, Synthetic minority oversampling technique.

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

Table 2 The effect of different machine learning algorithms on model prediction performance (Bootstrapping)

	AUC		Accuracy		Precision		Recall rate		F1 value	
	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI
machine learning algorithms										
AdaBoost	0.702±0.104	0.700-0.703	0.761±0.061	0.760-0.762	0.434±0.134	0.432-0.436	0.538±0.442	0.535-0.540	0.465±0.105	0.463-0.467
Bagging	0.749±0.083	0.748-0.750	0.776±0.064	0.774-0.777	0.457±0.137	0.454-0.459	0.486±0.159	0.483-0.489	0.452±0.112	0.450-0.454
Bernoulli NB	0.718±0.099	0.716-0.720	0.771±0.056	0.770-0.772	0.444±0.133	0.442-0.447	0.541±0.141	0.538-0.543	0.475±0.109	0.474-0.477
DT	0.667±0.085	0.665-0.668	0.738±0.067	0.737-0.739	0.388±0.127	0.386-0.390	0.491±0.151	0.489-0.494	0.417±0.105	0.416-0.419
Ensemble Learning	**0.793±0.083**	0.791-0.794	**0.810±0.058**	0.809-0.811	**0.545±0.157**	0.543-0.548	**0.576±0.162**	0.573-0.579	**0.537±0.108**	0.535-0.539
ET	0.596±0.097	0.594-0.598	0.703±0.081	0.701-0.704	0.308±0.149	0.305-0.310	0.393±0.186	0.390-0.396	0.326±0.139	0.324-0.329
Gaussian NB	0.667±0.106	0.665-0.669	0.720±0.061	0.719-0.721	0.364±0.106	0.362-0.366	0.543±0.133	0.541-0.545	0.429±0.103	0.427-0.431
Gradient Boosting	0.718±0.100	0.716-0.720	0.783±0.060	0.782-0.784	0.487±0.161	0.484-0.490	0.524±0.144	0.521-0.526	0.481±0.105	0.479-0.483
KNN	0.655±0.101	0.654-0.657	0.741±0.086	0.740-0.743	0.394±0.262	0.389-0.399	0.355±0.217	0.351-0.359	0.316±0.166	0.313-0.319
LDA	0.724±0.097	0.722-0.725	0.770±0.065	0.769-0.772	0.457±0.149	0.454-0.459	0.561±0.141	0.558-0.564	0.487±0.110	0.485-0.489
LR	0.728±0.094	0.727-0.730	0.770±0.070	0.769-0.771	0.465±0.155	0.462-0.467	0.580±0.143	0.577-0.583	0.497±0.110	0.495-0.499

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

Multinomial NB	0.727±0.099	0.725-0.728	0.753±0.071	0.752-0.754	0.450±0.170	0.447-0.453	0.570±0.175	0.567-0.573	0.467±0.111	0.465-0.469
Passive Aggressive	0.686±0.094	0.684-0.688	0.701±0.087	0.699-0.703	0.358±0.119	0.355-0.360	0.558±0.156	0.555-0.560	0.421±0.107	0.419-0.423
QDA	0.660±0.115	0.658-0.662	0.774±0.057	0.773-0.775	0.428±0.178	0.425-0.431	0.436±0.188	0.433-0.440	0.411±0.152	0.408-0.413
RF	0.742±0.088	0.741-0.744	0.792±0.075	0.791-0.793	0.534±0.194	0.531-0.538	0.430±0.175	0.427-0.432	0.444±0.119	0.441-0.446
SGD	0.720±0.099	0.718-0.722	0.762±0.064	0.761-0.764	0.452±0.196	0.448-0.455	0.507±0.213	0.503-0.511	0.434±0.141	0.432-0.437
SVM	0.735±0.090	0.734-0.737	0.792±0.073	0.790-0.793	0.533±0.194	0.529-0.536	0.443±0.165	0.440-0.446	0.449±0.115	0.447-0.451
XGBoost	0.740±0.095	0.738-0.741	0.790±0.074	0.789-0.792	0.515±0.161	0.512-0.518	0.513±0.165	0.510-0.516	0.486±0.112	0.484-0.488
p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	

Bernoulli NB, Bernoulli Naïve Bayes; DT, Decision Tree; ET, Extra Tree; Gaussian NB, Gaussian Naïve Bayes; KNN, K-Nearest Neighbor; LDA, Latent Dirichlet Allocation; LR, Logistic Regression; Multinomial NB, Multinomial Naïve Bayes; QDA, Quadratic Discriminant Analysis; SGD, Stochastic Gradient Descent; SVM, support vector machine. XGBoost, eXtreme Gradient Boosting.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

Table 3 Predictive performance indicators of the 5 best models

	AUC	accuracy	precision	recall rate	F1 value
model 1	**0.9141**	**0.8947**	**0.75**	0.6667	**0.7059**
model 2	0.9055	0.8105	0.5	**0.7778**	0.6087
model 3	0.9019	0.8421	0.6154	0.4444	0.5161
model 4	0.8997	0.8632	0.6316	0.6667	0.6486
model 5	0.8968	0.8316	0.5357	0.8333	0.6522

DISCUSSION

Traditional Chinese medicine has been used for the prevention and treatment of diseases for centuries [11]. In recent years, the application of Chinese herbal injections containing Panax notoginseng saponin has become more and more common in clinical practice, while ADR often causes concerns. Studies have shown that the Chinese herbal ingredients, traditional Chinese medicine preparation and combination medication are the important factors for the ADR of Chinese herbal injections containing Panax notoginseng saponin. Drug eruption (50.5%), allergic reactions (20.4%) and anaphylactic shock (9.7%) were the most common, and some cases were even life-threatening [5]. However, the ADR monitoring methods, including spontaneous reporting systems, prescription event monitoring and centralized hospital monitoring system, were all reported after the event, and may even have data bias, underreporting or repeated reporting. Therefore, the realization of ADR prediction has important significance for preventing ADR of Chinese herbal injections containing Panax notoginseng saponin in clinical practice.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

In our study, a nested case-control study was performed for data collection. In order to obtain the best model, we used 4 data filling, 5 data sampling and 3 variable selection methods for data processing, and combined 18 machine learning algorithms to establish 1080 ADR prediction models. By comparing the AUC, accuracy, precision, recall rate and F1 value of these models, the best one was selected to develop an ADR prediction system for the Chinese herbal injections containing Panax notoginseng saponin.

In recent years, some ADR prediction models have been developed based on data mining [6-9], machine learning algorithms [10, 12-15], and statistical methods [16-18]. Tangiisuran et al. [16] combined univariate analysis and multivariate binary logistic regression for the identification of clinical risk factors to develop an ADR risk model. The AUC of the model at the internal and external validation stage was 0.74 and 0.73, respectively, the sensitivity was 80% and 84%, and the specificity was 55% and 43% [16]. Imai et al. [10] used artificial neural networks to predict the ADR risk and made an AUC of 0.83. Compared with other studies, the model established in our study had better predictive performance (accuracy was 0.8947, precision was 0.75, the recall rate was 0.6667 and AUC was 0.914). As missing data is common in clinical practice, the methods of data filling used in our study may be advantageous for the deal with imbalanced data in clinical real-world research. More importantly, the system developed by the best model was potentially convenient for clinical application because of its' simple operation, fast calculation, and high accuracy.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

It is worth noting that Hammann et al. [19] established a decision tree model based on the chemical, physical, and structural properties of compounds for the prediction of ADR occurrence and the model had high predictive accuracy (78.9–90.2%). However, the model was difficult to interpret as it ignored the effect of pathological and physiological conditions and the combination medication on ADR. This made the model unlikely to be accepted by clinicians. In our study, we collected more than 80 factors including the patient's pathophysiological characteristics, clinical laboratory results, and medication conditions. Meanwhile, the critical predictors associated with the ADR were identified by the SHAP values. Although using the SHAP values as a generalized approach to identify the important clinical determinants of ADR caused by Chinese herbal injections containing Panax notoginseng saponin is not possible, it may help generate clinical hypotheses for some specific clinical events.

The results of SHAP indicated that whether the patients have dermatoses will significantly affect the models' predictive performance. Cutaneous ADR is one of the most common adverse reactions of Panax notoginseng, such as erythema multiforme, urticaria, severe erythema multiforme and acute generalized exanthematous pustulosis [20, 21]. Therefore, those patients with original dermatoses are more likely to have ADR after using Panax notoginseng. In addition, we found that age and gender are related to the occurrence of Panax notoginseng-induced ADR, which is consistent with the results reported by Yang et al. [22].

This study had some limitations. First, the small sample size of this study might

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

affect the model prediction performance. Second, as the study population was all from southwest China, the results may be biased while the prediction system was applied in other medical institutions. Finally, a prospective controlled trial is required to demonstrate the accuracy of the ADR prediction system.

Contributors XWW, EWL and RST were involved in the conception and design of the study. XWW drafted the article. JYZ, HC, XWS and YLW analyzed the data. EWL and RST revised the manuscript. All authors gave final approval of the version to be published. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. RST is the guarantor.

Funding This study was funded by the National Natural Science Foundation of China (No. 72004020), the Key Research and Development Program of Science and Technology Department of Sichuan Province (No. 2019YFS0514), the Postgraduate Research and Teaching Reform Project of the University of Electronic Science and Technology of China (No. JYJG201919) and the Research Subject of Health Commission of Sichuan Province (No. 19PJ262).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Ethical approval: This study was approved by the Ethics Committee of Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital (2017-11-01).

Provenance and peer review Not commissioned; externally peer reviewed.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Data availability statement Data are available upon reasonable request. Data may be obtained from a third party and are not publicly available. The first author (7190175@uestc.edu.cn) will share any publicly available data if requested by email.

Supplementary material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

REFERENCES

1 Xie W, Meng X, Zhai Y ,et al. Panax Notoginseng Saponins: A Review of Its Mechanisms of Antidepressant or Anxiolytic Effects and Network Analysis on Phytochemistry and Pharmacology. *Molecules* 2018; *23*.

2 Kim JH. Pharmacological and medical applications of Panax ginseng and

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

ginsenosides: a review for use in cardiovascular diseases. *J Ginseng Res* 2018; *42*:264-269.

3 Yang F, Ma Q, Matsabisa MG ,et al. Panax notoginseng for Cerebral Ischemia: A Systematic Review. *Am J Chin Med* 2020; *48*:1331-1351.

4 Qu J, Xu N, Zhang J ,et al. Panax notoginseng saponins and their applications in nervous system disorders: a narrative review. *Ann Transl Med* 2020; *8*:1525.

5 Xiang Z, Qiao T, Xiao H ,et al. The anaphylactoid constituents in Xue-Sai-Tong injection. *Planta Med* 2013; *79*:1043-1050.

6 Wei L, Scott J. Association rule mining in the US Vaccine Adverse Event Reporting System (VAERS). *Pharmacoepidemiol Drug Saf* 2015; *24*:922-933.

7 Harpaz R, DuMouchel W, Shah NH ,et al. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin Pharmacol Ther* 2012; *91*:1010-1021.

8 Sakaeda T, Tamon A, Kadoyama K ,et al. Data mining of the public version of the FDA Adverse Event Reporting System. *Int J Med Sci* 2013; *10*:796-803.

9 Kadoyama K, Kuwahara A, Yamamori M ,et al. Hypersensitivity reactions to anticancer agents: data mining of the public version of the FDA adverse event reporting system, AERS. *J Exp Clin Cancer Res* 2011; *30*:93.

10 Imai S, Takekuma Y, Kashiwagi H ,et al. Validation of the usefulness of artificial neural networks for risk prediction of adverse drug reactions used for

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

individual patients in clinical practice. *PLoS One* 2020; *15*:e0236789.

11 Liu SH, Chuang WC, Lam W ,et al. Safety surveillance of traditional Chinese medicine: current and future. *Drug Saf* 2015; *38*:117-128.

12 Choudhury O, Park Y, Salonidis T ,et al. Predicting Adverse Drug Reactions on Distributed Health Data using Federated Learning. *AMIA Annu Symp Proc* 2019; *2019*:313-322.

13 Liu X, Chen H. A research framework for pharmacovigilance in health social media: Identification and evaluation of patient adverse drug event reports. *J Biomed Inform* 2015; *58*:268-279.

14 Davis J, Costa VS, Peissig P ,et al. Demand-Driven Clustering in Relational Domains for Predicting Adverse Drug Events. *Proc Int Conf Mach Learn* 2012; *2012*:1287-1294.

15 Lee CY, Chen YP. Prediction of drug adverse events using deep learning in pharmaceutical discovery. *Brief Bioinform* 2021; *22*:1884-1901.

16 Tangiisuran B, Scutt G, Stevenson J ,et al. Development and validation of a risk model for predicting adverse drug reactions in older people during hospital stay: Brighton Adverse Drug Reactions Risk (BADRI) model. *PLoS One* 2014; *9*:e111254.

17 Clothier HJ, Lawrie J, Lewis G ,et al. SAEFVIC: Surveillance of adverse events following immunisation (AEFI) in Victoria, Australia, 2018. *Commun Dis Intell (2018)* 2020; *44*.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

18 Alvarez Y, Hidalgo A, Maignen F ,et al. Validation of statistical signal detection procedures in eudravigilance post-authorization data: a retrospective evaluation of the potential for earlier signalling. *Drug Saf* 2010; *33*:475-487.

19 Hammann F, Gutmann H, Vogt N ,et al. Prediction of adverse drug reactions using decision tree modeling. *Clin Pharmacol Ther* 2010; *88*:52-59.

20 Yan S, Xiong H, Shao F ,et al. HLA-C*12:02 is strongly associated with Xuesaitong-induced cutaneous adverse drug reactions. *Pharmacogenomics J* 2019; *19*:277-285.

21 Chen WJ, Kuang YY, Li JT. Analysis on 13 Cases of Adverse Drug Reaction by Xuesaitong Injection. *Journal of North Pharmacy* 2013; *10*:16-17.

22 Yang P, Qian N, Yao D ,et al. 62 Cases of Adverse Reactions in Xuesaitong Oral Preparations. *Chinese Medicine Modern Distance Education of China* 2021; *19*:34-36.

Figure 1 ROC curve of the 5 best models.

Figure 2 Importance matrix plot of each variable to the final prediction model. Variable names were shown in Supplementary Table 2. X83, pre-treatment serum levels; X55, renal function; X25, dermatoses; X1, gender; X2, age; X29, dose; X62, low-density lipoprotein; X64, hypoproteinemia; X30, anti-infective agents; X82, pre-treatment indicators of carcinoma; X79, hemoglobin; X6, history of allergy; X16, respiratory diseases; X66, albumin/globulin; X78, red blood cell; X81, hypersensitive

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

C-reactive protein; X51, dermatology medication; X77, eosinophils; X13, Charlson comorbidity index (Score); X57, serum potassium.

Figure 3 SHAP summary plot of the top 20 variables of the model. Red represents higher variable values, and blue represents lower variable values. Variable names were shown in Supplementary Table 2. X83, pre-treatment serum levels; X55, renal function; X25, dermatoses; X1, gender; X2, age; X29, dose; X62, low-density lipoprotein; X64, hypoproteinemia; X30, anti-infective agents; X82, pre-treatment indicators of carcinoma; X79, hemoglobin; X6, history of allergy; X16, respiratory diseases; X66, albumin/globulin; X78, red blood cell; X81, hypersensitive C-reactive protein; X51, dermatology medication; X77, eosinophils; X13, Charlson comorbidity index (Score); X57, serum potassium.

Figure 4 Sample size validation. The vertical bars represent the 95% confidence interval (CI) of AUC of ROC.

Figure 5 The development of ADR prediction system.

Figure 6 The operation (A) and output (B) of the ADR prediction system.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

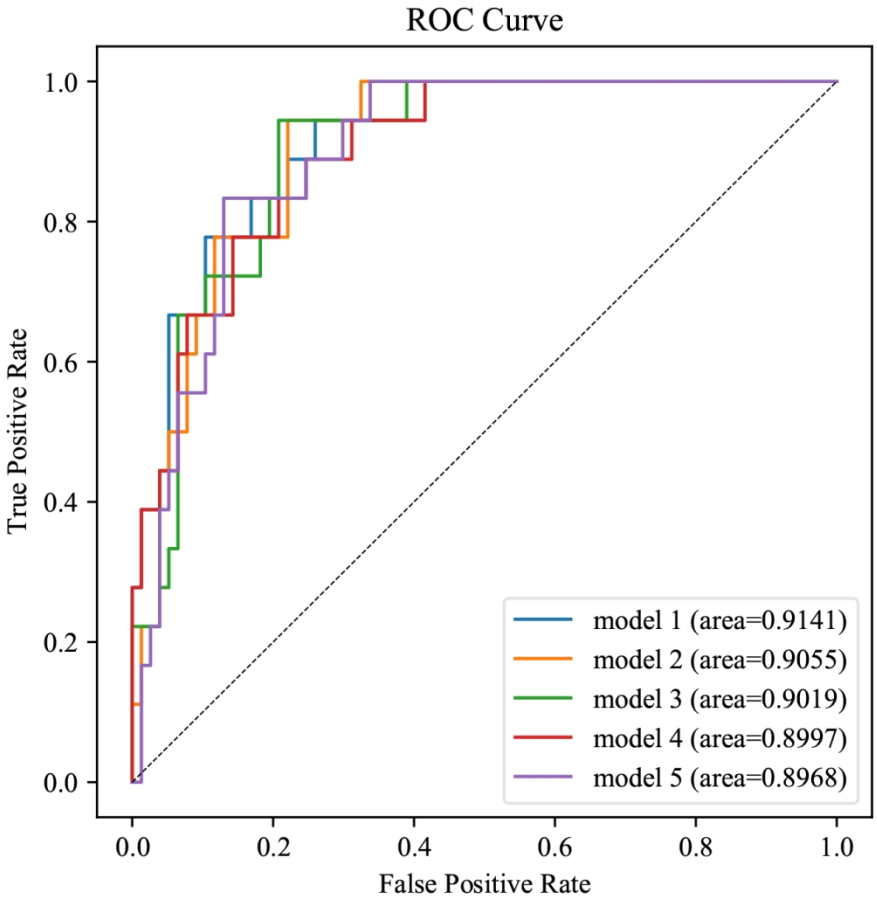

Figure 1 ROC curve of the 5 best models.

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

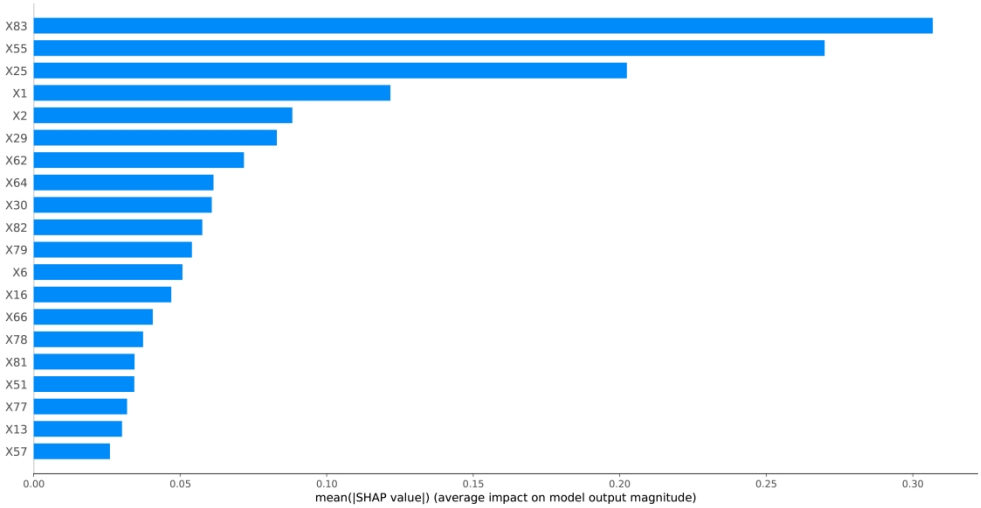

Figure 2 Importance matrix plot of each variable to the final prediction model.

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

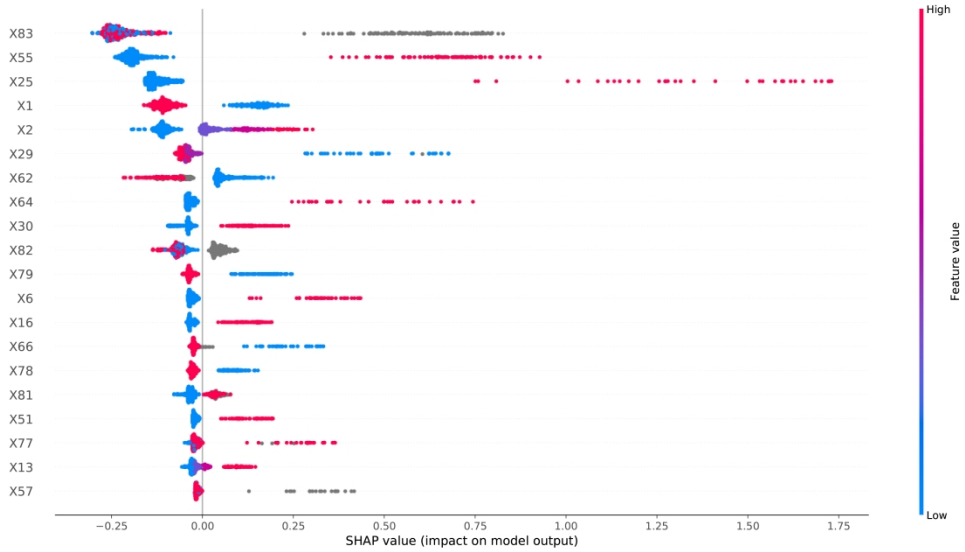

Figure 3 SHAP summary plot of the top 20 variables of the model.

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

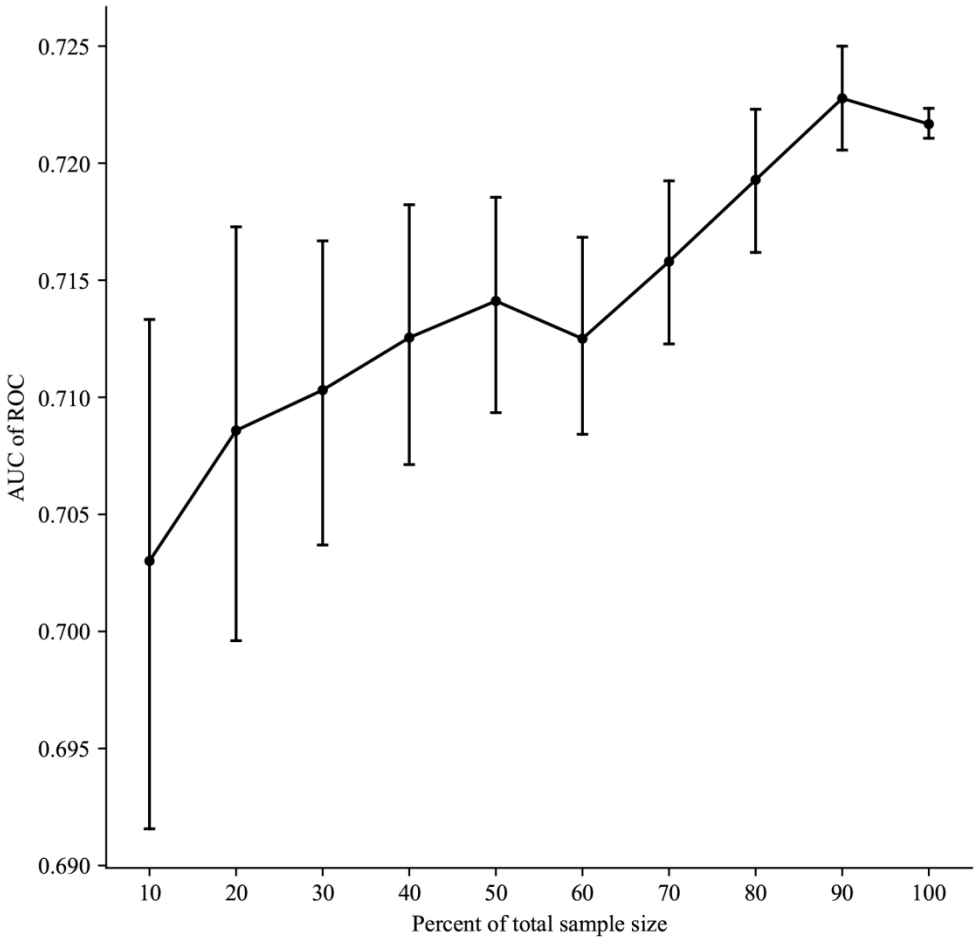

Figure 4 Sample size validation.

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

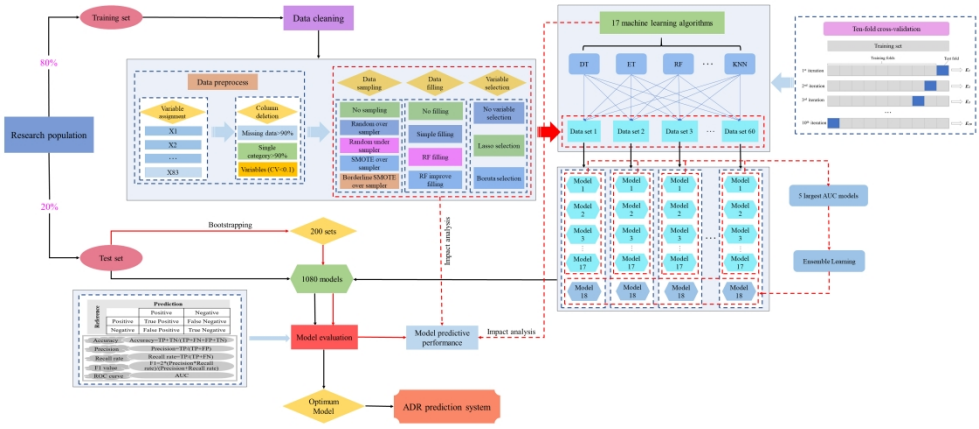

Figure 5 The development of ADR prediction system.

1050x472mm (96 x 96 DPI)

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

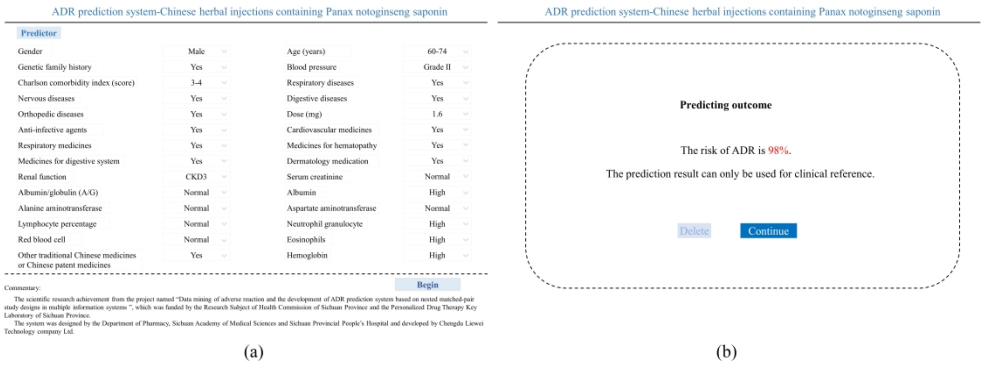

Figure 6 The operation (A) and output (B) of the ADR prediction system.

1899x688mm (96 x 96 DPI)

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

Table 1 Demographic and clinical characteristics of the patients

Parameter	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 (%)

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Table 2 Variable assignment

Number	Variable	Assignment
	Adverse drug reaction	1, Yes; 0, No
X1	Gender	1, Male; 0, Female
X2	Age (years)	1, ≤ 44; 2, 45 ≤ Age ≤ 59; 3, 60 ≤ Age ≤ 74; 4, ≥ 75
X3	Body mass index (BMI, kg/m^2)	1, < 18.5; 2, 18.5 ≤ BMI ≤ 23.9; 3, ≥ 24
X4	Asians	1, Yes; 0, No
X5	Genetic family history	1, Yes; 0, No
X6	History of allergy	1, Yes; 0, No
X7	Smoking	1, Yes; 0, No
X8	Alcohol	1, Yes; 0, No
X9	Temperature (℃)	1, < 36.1; 2, 36.1 ≤ Temperature ≤ 37.2; 3, > 37.3
X10	Pulse (beats/min)	1, < 60; 2, 60 ≤ Pulse ≤ 100, 3, > 100
X11	Breathe (times/min)	1, < 12; 2, 12 ≤ Breathe ≤ 20; 3, > 20
X12	Blood pressure	0, Normal (systolic pressure ≤ 139 mmHg or diastolic pressure ≤ 89 mmHg); 1, Grade I (140 mmHg ≤ systolic pressure ≤ 159 mmHg or 90 mmHg ≤ diastolic pressure ≤ 99 mmHg); 2, Grade II (160 mmHg ≤ systolic pressure ≤ 179 mmHg or 100 mmHg ≤ diastolic pressure ≤ 109 mmHg); 3, Grade III (systolic pressure ≥180 mmHg or diastolic pressure ≥110 mmHg)
X13	Charlson comorbidity index (Score)	1, 0; 2, 1 or 2; 3, 3 or 4; 4, ≥ 5
X14	Cardiovascular disease	1, Yes; 0, No

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

X15	Endocrine diseases	1, Yes; 0, No
X16	Respiratory diseases	1, Yes; 0, No
X17	Nervous diseases	1, Yes; 0, No
X18	Digestive diseases	1, Yes; 0, No
X19	Neoplastic diseases	1, Yes; 0, No
X20	Orthopedic diseases	1, Yes; 0, No
X21	Genito-urinary diseases	1, Yes; 0, No
X22	Hematopathy	1, Yes; 0, No
X23	Oculopathy	1, Yes; 0, No
X24	Ear-nose-throat diseases	1, Yes; 0, No
X25	Dermatoses	1, Yes; 0, No
X26	Immune rheumatism	1, Yes; 0, No
X27	Other diseases	1, Yes; 0, No
X28	Solvent	1, 0.9% sodium chloride injection; 2, 5% glucose injection; 3, Other solvents
X29	Dose (mg)	1, < 1.6; 2, =1.6; 3, > 1.6
X30	Anti-infective agents	1, Yes; 0, No
X31	Cardiovascular medicines	1, Yes; 0, No
X32	Medicines for digestive system	1, Yes; 0, No
X33	Respiratory medicines	1, Yes; 0, No
X34	Nervous system medicines	1, Yes; 0, No
X35	Medication in mental disorders	1, Yes; 0, No

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

X36	Non-steroidal anti-inflammatory drugs	1, Yes; 0, No
X37	Antiallergic agent	1, Yes; 0, No
X38	Genito-urinary system medicines	1, Yes; 0, No
X39	Medicines for hematopathy	1, Yes; 0, No
X40	Endocrine agents or hormone drugs	1, Yes; 0, No
X41	Antineoplastic drugs	1, Yes; 0, No
X42	Amino acids, vitamins, minerals or other nutrition preparations	1, Yes; 0, No
X43	Regulating water, electrolyte or acid-base balance drugs	1, Yes; 0, No
X44	Adjuvant agents to anesthesia or anesthetics	1, Yes; 0, No
X45	Diagnostic agents	1, Yes; 0, No
X46	Biological agents	1, Yes; 0, No
X47	Obstetrical-gynecological drugs	1, Yes; 0, No
X48	Stomatological preparations	1, Yes; 0, No
X49	Ophthalmic medication	1, Yes; 0, No
X50	Ear-nose-throat medication	1, Yes; 0, No
X51	Dermatology medication	1, Yes; 0, No
X52	Other traditional Chinese medicines	1, Yes; 0, No

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	or Chinese patent medicines	
X53	Urea	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X54	Serum creatinine	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X55	Renal function	1, Glomerular filtration rate ≥ 90 ml/(min·1.73m^2); 2, 60ml/(min·1.73m^2) ≤ Glomerular filtration rate ≤ 89ml/(min·1.73m^2); 3, 30ml/(min·1.73m^2) ≤ Glomerular filtration rate ≤59 ml/(min·1.73m^2); 4, 15ml/(min·1.73m^2) ≤ Glomerular filtration rate ≤29 ml/(min·1.73m^2); 5, Glomerular filtration rate < 15 ml/(min·1.73m^2)
X56	Blood glucose	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X57	Serum potassium	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X58	Serum sodium	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X59	Total cholesterol	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X60	Triglyceride	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X61	High-density lipoprotein	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X62	Low-density lipoprotein	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X63	Albumin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X64	Hypoproteinemia	1, Yes; 0, No
X65	Globulin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X66	Albumin/globulin (A/G)	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X67	Aspartate aminotransferase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X68	Alanine aminotransferase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

X69	Liver function	1, Less than 3 times upper limit of normal range of liver function tests (ULN of LFTs); 2, 3~5 times ULN of LFTs; 3, More than 5 times ULN of LFTs
X70	Total bilirubin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X71	Lactic dehydrogenase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X72	Creatine kinase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X73	White blood cell	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X74	Neutrophil granulocyte	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X75	Lymphocyte percentage	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X76	Monocyte percentage	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X77	Eosinophils	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X78	Red blood cell	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X79	Hemoglobin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X80	Platelet count	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
X81	Hypersensitive C-reactive protein	0, Within the normal range; 1, Above the normal range
X82	Pre-treatment indicators of carcinoma	0, Within the normal range; 1, Above the normal range
X83	Pre-treatment serum levels	0, Within the normal range; 1, Above the normal range

bmjopen-2022-061457 on 18 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Table 3 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.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Table 4 The effect of different data processing methods and machine learning algorithms on model prediction performance (Ten-fold cross-validation)

		AUC		Accuracy		Precision		Recall rate		F1 value	
		Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI	Mean±SD	95%CI
Data filling											
	No filling	0.868±0.099	0.864-0.872	0.820±0.093	0.816-0.823	0.772±0.190	0.765-0.779	0.720±0.254	0.710-0.730	0.729±0.217	0.721-0.737
	Simple filling	0.881±0.097	0.877-0.885	0.828±0.100	0.824-0.832	0.793±0.165	0.787-0.799	0.746±0.243	0.737-0.756	0.751±0.197	0.744-0.759
	RF filling	0.885±0.095	0.881-0.888	0.831±0.095	0.827-0.835	**0.802±0.157**	0.796-0.808	0.749±0.237	0.740-0.759	**0.757±0.189**	0.750-0.764
	RF improve filling	**0.887±0.094**	0.883-0.890	**0.832±0.096**	0.828-0.835	0.799±0.158	0.793-0.806	**0.751±0.240**	0.742-0.760	0.757±0.191	0.749-0.764
	p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	
Data sampling											
	No sampling	0.824±0.088	0.820-0.828	0.832±0.050	0.830-0.835	0.641±0.271	0.629-0.653	0.399±0.197	0.391-0.408	0.464±0.193	0.455-0.472
	Random over sampler	**0.923±0.063**	0.920-0.925	0.858±0.085	0.854-0.861	**0.849±0.079**	0.845-0.852	0.872±0.118	0.867-0.877	0.857±0.089	0.854-0.861
	Random under sampler	0.815±0.107	0.810-0.819	0.732±0.104	0.728-0.737	0.783±0.145	0.776-0.789	0.678±0.188	0.670-0.686	0.707±0.132	0.701-0.713
	SMOTE over sampler	0.920±0.072	0.917-0.923	0.857±0.081	0.853-0.860	0.844±0.071	0.841-0.848	0.875±0.125	0.869-0.880	0.856±0.089	0.852-0.860
	Borderline SMOTE	0.919±0.077	0.916-0.923	**0.859±0.085**	0.855-0.862	0.841±0.074	0.837-0.844	**0.885±0.130**	0.879-0.890	**0.859±0.093**	0.855-0.863
	p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	
Variable selection											

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	No selection	0.870±0.105	0.867-0.874	0.820±0.104	0.817-0.824	0.780±0.178	0.774-0.786	0.733±0.254	0.725-0.742	0.737±0.208	0.730-0.744
	Lasso selection	**0.889±0.089**	0.886-0.892	**0.835±0.090**	0.832-0.838	**0.801±0.165**	0.796-0.807	**0.751±0.240**	0.743-0.759	**0.758±0.196**	0.752-0.765
	Boruta selection	0.881±0.094	0.878-0.884	0.827±0.093	0.824-0.830	0.794±0.162	0.788-0.799	0.741±0.236	0.733-0.749	0.750±0.191	0.744-0.757
	p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	
machine learning algorithms											
	AdaBoost	0.871±0.092	0.864-0.879	0.813±0.093	0.806-0.820	0.784±0.136	0.773-0.795	0.731±0.202	0.715-0.747	0.745±0.160	0.733-0.758
	Bagging	0.907±0.102	0.898-0.915	0.854±0.101	0.846-0.863	0.805±0.158	0.793-0.818	0.791±0.245	0.771-0.810	0.785±0.196	0.769-0.801
	Bernoulli NB	0.866±0.082	0.860-0.873	0.802±0.085	0.795-0.809	0.771±0.144	0.759-0.783	0.719±0.178	0.705-0.733	0.736±0.148	0.724-0.748
	DT	0.815±0.110	0.806-0.824	0.805±0.089	0.797-0.812	0.773±0.158	0.760-0.786	0.715±0.237	0.696-0.734	0.724±0.184	0.709-0.739
	ET	0.829±0.110	0.821-0.838	0.809±0.092	0.801-0.816	0.767±0.164	0.754-0.780	0.714±0.255	0.694-0.735	0.720±0.207	0.704-0.737
	Gaussian NB	0.845±0.089	0.838-0.852	0.786±0.085	0.779-0.793	0.734±0.155	0.722-0.747	0.743±0.164	0.730-0.756	0.730±0.143	0.719-0.742
	Gradient Boosting	0.891±0.102	0.883-0.899	0.841±0.099	0.833-0.849	0.822±0.149	0.810-0.834	0.746±0.252	0.725-0.766	0.762±0.194	0.747-0.778
	KNN	0.896±0.084	0.890-0.903	0.830±0.098	0.822-0.838	0.747±0.296	0.724-0.771	0.687±0.381	0.656-0.717	0.674±0.326	0.648-0.700
	LDA	0.897±0.073	0.891-0.903	0.835±0.081	0.829-0.842	0.805±0.117	0.796-0.815	0.768±0.191	0.753-0.783	0.777±0.144	0.765-0.788
	LR	0.893±0.076	0.886-0.899	0.834±0.082	0.827-0.840	0.815±0.119	0.805-0.824	0.754±0.216	0.737-0.772	0.767±0.157	0.755-0.780
	Multinomial NB	0.839±0.071	0.834-0.845	0.773±0.078	0.766-0.779	0.753±0.161	0.740-0.766	0.653±0.235	0.634-0.672	0.676±0.190	0.660-0.691
	Passive Aggressive	0.836±0.098	0.828-0.844	0.780±0.091	0.772-0.787	0.723±0.161	0.711-0.736	0.720±0.205	0.703-0.736	0.712±0.172	0.698-0.725

For peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

QDA	0.915±0.081	0.909-0.922	0.860±0.089	0.853-0.868	0.827±0.152	0.814-0.839	0.798±0.184	0.783-0.812	0.805±0.156	0.792-0.817
RF	0.919±0.097	0.911-0.926	0.871±0.100	0.863-0.879	0.843±0.154	0.831-0.856	0.775±0.268	0.753-0.796	0.788±0.214	0.771-0.805
SGD	0.895±0.075	0.889-0.901	0.832±0.082	0.825-0.839	0.803±0.197	0.787-0.819	0.710±0.287	0.687-0.733	0.726±0.238	0.707-0.745
SVM	**0.926±0.086**	0.919-0.933	**0.875±0.096**	0.867-0.883	**0.858±0.144**	0.847-0.870	0.776±0.271	0.754-0.797	0.791±0.217	0.773-0.808
XGBoost	0.922±0.092	0.914-0.929	0.869±0.100	0.861-0.877	0.825±0.153	0.812-0.837	**0.810±0.229**	0.792-0.828	**0.808±0.185**	0.793-0.822
p value	***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**		***p*<0.0001**	

AUC, Area under curve; RF, Random Forest; SMOTE, Synthetic minority oversampling technique; Bernoulli NB, Bernoulli Naïve Bayes; DT, Decision Tree; ET, Extra Tree; Gaussian NB, Gaussian Naïve Bayes; KNN, K-Nearest Neighbor; LDA, Latent Dirichlet Allocation; LR, Logistic Regression; Multinomial NB, Multinomial Naïve Bayes; QDA, Quadratic Discriminant Analysis; SGD, Stochastic Gradient Descent; SVM, support vector machine. XGBoost, eXtreme Gradient Boosting

For Peer review only

bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Reporting checklist for prediction model development/validation.

Based on the TRIPOD guidelines.

Instructions to authors

Complete this checklist by entering the page numbers from your manuscript where readers will find each of the items listed below.

Your article may not currently address all the items on the checklist. Please modify your text to include the missing information. If you are certain that an item does not apply, please write "n/a" and provide a short explanation.

Upload your completed checklist as an extra file when you submit to a journal.

In your methods section, say that you used the TRIPODreporting guidelines, and cite them as:

Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement.

		Reporting Item	Page Number
Title			
	#1	Identify the study as developing and / or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract			

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only

	#2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2
Introduction			
	#3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	3
	#3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	4
Methods			
Source of data	#4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	5
Source of data	#4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	5
Participants	#5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5
Participants	#5b	Describe eligibility criteria for participants.	5
Participants	#5c	Give details of treatments received, if relevant	5

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

Outcome	#6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	7
Outcome	#6b	Report any actions to blind assessment of the outcome to be predicted.	7
Predictors	#7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured	6
Predictors	#7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	6
Sample size	#8	Explain how the study size was arrived at.	5
Missing data	#9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	6
Statistical analysis methods	#10a	If you are developing a prediction model describe how predictors were handled in the analyses.	6
Statistical analysis methods	#10b	If you are developing a prediction model, specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	7
Statistical analysis methods	#10c	If you are validating a prediction model, describe how the predictions were calculated.	7
Statistical analysis methods	#10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	7

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only

Statistical analysis methods	#10e	If you are validating a prediction model, describe any model updating (e.g., recalibration) arising from the validation, if done	7
Risk groups	#11	Provide details on how risk groups were created, if done.	7
Development vs. validation	#12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	7
Results			
Participants	#13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	8
Participants	#13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	8
Participants	#13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	8
Model development	#14a	If developing a model, specify the number of participants and outcome events in each analysis.	9
Model development	#14b	If developing a model, report the unadjusted association, if calculated between each candidate predictor and outcome.	9

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

Model specification	#15a	If developing a model, present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	9
Model specification	#15b	If developing a prediction model, explain how to the use it.	9
Model performance	#16	Report performance measures (with CIs) for the prediction model.	9
Model-updating	#17	If validating a model, report the results from any model updating, if done (i.e., model specification, model performance).	9
Discussion			
Limitations	#18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	17
Interpretation	#19a	For validation, discuss the results with reference to performance in the development data, and any other validation data	15
Interpretation	#19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	15
Implications	#20	Discuss the potential clinical use of the model and implications for future research	16

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.

For peer review only

Other information

Supplementary information	#21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	18
Funding	#22	Give the source of funding and the role of the funders for the present study.	18

None The TRIPOD checklist is distributed under the terms of the Creative Commons Attribution License CC-BY. This checklist can be completed online using https://www.goodreports.org/, a tool made by the EQUATOR Network in collaboration with Penelope.ai

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 8 September 2022. Downloaded from http://bmjopen.bmj.com/ on April 26, 2024 by guest. Protected by copyright.