

# 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](mailto:info.bmjopen@bmj.com)

# BMJ Open

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

Journal:	<i>BMJ Open</i>
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



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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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

4     Xing-Wei Wu<sup>1,2</sup>, Jia-Ying Zhang<sup>3</sup>, Huan Chang<sup>1</sup>, Xue-Wu Song<sup>1,2</sup>, Ya-Lin Wen<sup>1</sup>, En-  
5     Wu Long<sup>1,2</sup>, Rong-Sheng Tong<sup>1,2</sup>

6     <sup>1</sup>Personalized Drug Therapy Key Laboratory of Sichuan Province, School of Medicine,  
7     University of Electronic Science and Technology of China, Chengdu 610072, China,

8     <sup>2</sup>Department of Pharmacy, Sichuan Academy of Medical Sciences and Sichuan  
9     Provincial People's Hospital, Chengdu 610072, China,

10    <sup>3</sup>Department of Pharmacy, Chengdu First People's Hospital, Chengdu 610095, China

11    **Correspondence to**

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

16    [318004031@qq.com](mailto:318004031@qq.com)

17    **Word count: 2225**

1  
2  
3  
4 18 **Develop an ADR prediction system of Chinese herbal injection**  
5  
6 19 **containing Panax notoginseng saponin: a nested case-control study**  
7  
8  
9 20 **using machine learning**  
10  
11

12 21 **ABSTRACT**  
13  
14

15 22 **Objective** This study aimed to develop an adverse drug reactions (ADR) antecedent  
16  
17  
18 23 prediction system using machine learning algorithms to provide the reference for  
19  
20  
21 24 security usage of Chinese herbal injections containing Panax notoginseng saponin in  
22  
23  
24 25 clinical practice.

25  
26 26 **Design** A nested case-control study.  
27

28 27 **Setting** National Center for ADR Monitoring and the Electronic Medical Record (EMR)  
29  
30  
31 28 system.  
32

33 29 **Participants** All patients were from 5 medical institutions in Sichuan Province from  
34  
35  
36 30 January 2010 to December 2018.  
37

38  
39 31 **Main outcomes/measures** Information of patients with ADR who using Chinese  
40  
41  
42 32 herbal injections containing Panax notoginseng saponin was collected from the  
43  
44  
45 33 National Center for ADR Monitoring. A nested case-control study was used to  
46  
47  
48 34 randomly match patients without ADR from the EMR system according to 1:4.  
49  
50  
51 35 Eighteen machine learning algorithms were applied for the development of ADR  
52  
53  
54 36 prediction models. Area under curve (AUC), accuracy, precision, recall rate and F1  
55  
56  
57 37 value were used to evaluate the predictive performance of the model. An ADR  
58  
59  
60 38 prediction system were established by the optimal model selected from the 1080 models.

39 39 **Results** A total of 530 patients from 5 medical institutions were included, and 1080

1  
2  
3  
4 40 ADR prediction models were developed. Among these models, the AUC of the best  
5  
6 41 capable one was 0.9141 and the accuracy was 0.8947. According to the parameters of  
7  
8  
9 42 the best model, a prediction system for the ADR of Panax notoginseng saponin has been  
10  
11 43 established, which can realize the output of patient ADR risk.

12  
13  
14 44 **Conclusion** The prediction system developed based on the machine learning model in  
15  
16  
17 45 this study had good predictive performance and potential clinical application.

18  
19 46 **Key words** Adverse drug reactions, Chinese herbal injection, Machine learning,  
20  
21  
22 47 Prediction system, Panax notoginseng saponin

#### 23 24 25 48 **Strengths and limitations of this study**

26  
27 49 ➤ We first used machine learning to predict the ADR of Chinese herbal injection  
28  
29  
30 50 containing Panax notoginseng saponin.

31  
32 51 ➤ Eighteen machine learning algorithms were used to establish 1080 ADR prediction  
33  
34  
35 52 models. An ADR prediction system with Chinese herbal injections containing  
36  
37  
38 53 Panax notoginseng saponin developed by the best model had high accuracy and  
39  
40  
41 54 precision, and had potential value for clinical application.

42  
43 55 ➤ More than 80 factors including the patient's pathophysiological characteristics,  
44  
45  
46 56 clinical laboratory results, and medication conditions, were incorporated in our  
47  
48  
49 57 study.

50  
51 58 ➤ More data were needed to further evaluate the model prediction performance.

#### 52 53 54 59 **INTRODUCTION**

55  
56 60 Panax notoginseng saponins, as the main ingredients of Panax notoginseng (Buck.)

57  
58  
59 61 F.H.Chen, has been widely used in the disease therapy of nervous system and cardio-

1  
2  
3  
4 62 cerebral vascular system <sup>1-4</sup>. High frequency of adverse drug reactions (ADR) in  
5  
6 63 Chinese herbal containing Panax notoginseng saponin has received widespread  
7  
8  
9 64 attention. Of all the adverse reactions, about 69.57% were caused by injections, mainly  
10  
11 65 manifested as drug eruption (50.5%), allergic reaction (20.4%) and anaphylactic shock  
12  
13  
14 66 (9.7%), which can be life-threatening in severe cases <sup>5</sup>.

15  
16  
17 67 At present, ADR is mainly monitored by spontaneous reporting system, case-  
18  
19 68 control study, cohort study, prescription event monitoring and centralized hospital  
20  
21 69 monitoring system. However, most of these methods have obvious hysteresis.  
22  
23  
24 70 Therefore, there is an increasing need to develop an ADR antecedent prediction system  
25  
26  
27 71 to prevent and avoid the occurrence of ADR in Chinese herbal injections containing  
28  
29  
30 72 Panax notoginseng saponin.

31  
32 73 Machine learning, the core technology of artificial intelligence, is commonly used  
33  
34  
35 74 to build prediction models. In recent years, some prediction models for ADR have been  
36  
37  
38 75 established <sup>6-10</sup>. Based on a clustering method for the postprocessing of association rules,  
39  
40  
41 76 Lai et al. <sup>6</sup> developed an application of stepwise association rule mining to identify the  
42  
43  
44 77 associations between vaccine and multiple adverse events. In addition, Imai et al. <sup>10</sup>  
45  
46  
47 78 used artificial neural networks to evaluate vancomycin-induced nephrotoxicity.  
48  
49  
50 79 However, small sample size, incomplete patient information, and unsatisfactory  
51  
52  
53 80 predictive performance restrict the application of ADR prediction models in clinical  
54  
55  
56 81 practice. In view of these challenges, this study collected patients information in the  
57  
58  
59 82 National Center for ADR Monitoring and the Electronic Medical Record (EMR) system  
60  
61  
62 83 by a nested case-control study to establish an ADR prediction model of Chinese herbal

1  
2  
3  
4 84 injections containing Panax notoginseng saponin, and develop an ADR prediction  
5  
6 85 system based on machine learning algorithms to provide reference for clinical ADR  
7  
8  
9 86 management and prevention.  
10

## 11 87 **METHODS**

### 12 13 14 15 88 **Data collection**

16  
17  
18 89 Information of patients with ADR who using Chinese herbal injections containing  
19  
20 90 Panax notoginseng saponin from the National Center for ADR Monitoring was  
21  
22  
23 91 collected. A nested case-control study was used to randomly match patients without  
24  
25  
26 92 ADR who using Chinese herbal injections containing Panax notoginseng saponin from  
27  
28  
29 93 the EMR system according to 1:4. All patients were from 5 medical institutions in  
30  
31  
32 94 Sichuan Province from January 2010 to December 2018. This study was approved by  
33  
34 95 the Ethics Committee of Sichuan Academy of Medical Sciences and Sichuan Provincial  
35  
36  
37 96 People's Hospital.

### 38 39 40 97 **Data cleaning**

#### 41 42 98 *Variable assignment*

43  
44  
45 99 Binary-state variables were directly assigned values of 0 or 1. According to whether in  
46  
47  
48 100 the normal range, clinical laboratory variables were assigned values of 1, 2 and 3 (1,  
49  
50  
51 101 below the normal range; 2, within the normal range; and 3, above the normal range).

#### 52 53 102 *Column deletion*

54  
55  
56 103 Variables with missing data >90%, or a single category >90%, or the coefficient of  
57  
58  
59 104 variation (CV) <0.1 were deleted.  
60



1  
2  
3  
4 105 *Data filling*  
5

6 106 There are 4 ways to data filling. No filling means to retain the original data directly.  
7  
8

9 107 Simple filling refers to use the mean fill for continuous variables, the mode for  
10

11 108 disordered categorical variables, and the median for ordered categorical variables.  
12  
13

14 109 Random Forest (RF) filling orders the column according to the number of missing data,  
15

16 110 and then the missing data was predicted and filled by RF model. RF improve filling  
17

18 111 refers to predict and fill the column with the least missing data, which was used as the  
19  
20

21 112 input for the prediction and filling of other missing data.  
22  
23

24 113 *Data sampling*  
25

26 114 No sampling: directly input the original data into the model. Random over sampler:  
27  
28

29 115 random replication of data with fewer types to make the sample sizes of different types  
30  
31

32 116 consistent, while random under sampler is to randomly delete data with more types.  
33  
34

35 117 Synthetic minority oversampling technique (SMOTE) over sampler: synthesize new  
36  
37

38 118 data by analyzing a small amount of original data. Borderline SMOTE over sampler:  
39  
40

41 119 synthesize new data from borderline data.  
42

43 120 *Variable selection*  
44

45 121 No variable selection or use Lasso or Boruta for variable selection.  
46  
47

48 122 **Model establishment**  
49

50 123 Through different data filling, data sampling and variable selection, 60 data sets were  
51  
52

53 124 obtained. Eighteen machine learning algorithms, including AdaBoost, Bagging,  
54  
55

56 125 Bernoulli Naïve Bayes (Bernoulli NB), Decision Tree (DT), Extra Tree (ET), Gaussian  
57  
58

59 126 Naïve Bayes (Gaussian NB), Gradient Boosting, K-Nearest Neighbor (KNN), Latent  
60

1  
2  
3  
4 127 Dirichlet Allocation (LDA), Logistic Regression (LR), Multinomial Naïve Bayes  
5  
6 128 (Multinomial NB), Passive Aggressive, Quadratic Discriminant Analysis (QDA), RF,  
7  
8  
9 129 Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), eXtreme  
10  
11 130 Gradient Boosting (XGBoost), and Ensemble Learning, were used to build models.

12  
13  
14 131 The model establishment was as follows. The data was standardized and divided  
15  
16 132 into a training set and a test set according to 8:2. The training set was used to build  
17  
18 133 models, and the test set was used to evaluate the predictive performance of the models.  
19  
20 134 Ten-fold cross-validation on the training set was used for internal validation of the  
21  
22 135 model, and 200 Bootstrapping samples from the test set for the evaluation of the impact  
23  
24 136 of different data processing methods or machine learning algorithms on model  
25  
26 137 predictive performance. Five algorithms with the largest area under curve (AUC) on  
27  
28 138 each data set were used for ensemble learning.

### 139 **Model evaluation**

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

### 146 **Sample size assessment**

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

1  
2  
3  
4 148 randomly extracted 10%, 20%, 30% to 100% subsets from the training set by  
5  
6 149 Bootstrapping. The 10 subsets were used to establish models, respectively. Repeated  
7  
8  
9 150 the procedure 100 times and the AUC, calculated from the testing set, was used for  
10  
11  
12 151 sample size examination.

## 13 14 15 152 **Patient and public involvement**

16  
17  
18 153 Patients and/or the public were not directly involved in this study.

## 19 20 21 154 **Statistical Analysis**

22  
23  
24 155 Categorical variables were expressed as counts and percentages and continuous  
25  
26 156 variables as mean  $\pm$  standard deviation. Analysis of variance will be used if the data  
27  
28  
29 157 were normally distributed and the variances were equal, otherwise, Kruskal-Wallis test  
30  
31  
32 158 will be used.  $p$  value  $< 0.05$  were considered statistically significant. Hypothesis testing  
33  
34 159 and Models building were implemented using the stats and sklearn packages in Python  
35  
36  
37 160 (Version3.8), respectively.

## 38 39 40 161 **RESULTS**

### 41 42 43 162 **Research population**

44  
45  
46 163 A total of 530 patients were enrolled in this study, of which 106 patients had ADR.  
47  
48  
49 164 ADR patients included 50 (47.17%) males and 56 (52.83%) females.

### 50 51 52 165 **Data cleaning**

53  
54  
55 166 The assignment of all variables was shown in Supplementary Table 1. After data  
56  
57  
58 167 processing by 4 data filling, 5 data sampling and 3 variable selection methods, we  
59  
60

1  
2  
3  
4 168 obtained 60 data sets. The results of variable selection by the Lasso and Boruta were  
5  
6 169 shown in Supplementary Figure 1.  
7  
8  
9

### 10 170 **Model establishment**

11  
12 171 A total of 1080 prediction models were established by 18 machine learning algorithms  
13  
14 172 and the 60 data sets. The results of ten-fold cross-validation were shown in  
15  
16 173 Supplementary Table 2. Using 200 Bootstrapping samples from the test set to evaluate  
17  
18 174 the impact of different data processing methods or machine learning algorithms on  
19  
20 175 model predictive performance. The results showed that differences of model predictive  
21  
22 176 performance exist by different data filling, data sampling, variable selection (Table 1)  
23  
24 177 and machine learning algorithms (Table 2). The ensemble learning model had the best  
25  
26 178 performance with an AUC of  $0.793 \pm 0.083$  (Table 2).  
27  
28  
29  
30  
31  
32  
33

### 34 179 **Model evaluation**

35  
36  
37 180 The AUC, accuracy, precision, recall rate, and F1 value were used to evaluate the  
38  
39 181 performance of the model. The best 5 prediction models were selected and model 1 had  
40  
41 182 the best performance with an AUC of 0.9141 (Table 3). The receiver operating  
42  
43 183 characteristic (ROC) curve of the 5 best model was shown in Figure 1.  
44  
45  
46  
47

### 48 184 **Model interpretation**

49  
50  
51 185 The importance of each variable to the final prediction model was shown in Figure 2.  
52  
53 186 The result showed that pre-treatment serum levels, renal function, dermatoses, gender  
54  
55 187 and age were the top 5 most important variables contributing to the model. We used the  
56  
57 188 SHAP value to explain the contribution of the variables to the model, and the SHAP  
58  
59  
60

1  
2  
3  
4 189 value of the top 20 variables was shown in Figure 3. This plot explains how high and  
5  
6 190 low variables values were in relation to SHAP values. According to the prediction  
7  
8  
9 191 model, the higher the SHAP value of a variable, the more likely ADR occurs.  
10

## 11 12 192 **Sample size assessment**

13  
14  
15 193 With the continuous increased size of sample data, the AUC values of the testing sets  
16  
17  
18 194 continued to increase, which shows a sufficient sample size was included in this study  
19  
20  
21 195 (Figure 4).  
22

## 23 24 196 **Develop an ADR prediction system for Panax notoginseng saponin**

25  
26  
27 197 According to the parameters of the best model, a prediction system for the ADR of  
28  
29  
30 198 Panax notoginseng saponin has been developed and we had obtained the software  
31  
32  
33 199 copyright. The development of ADR prediction system was shown in Figure 5. The  
34  
35 200 operation and output of the system were shown in Figure 6.  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

201 **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	<b>0.786±0.101</b>	0.785-0.787	<b>0.770±0.070</b>	0.769-0.771	0.437±0.162	0.435-0.438	<b>0.546±0.208</b>	0.544-0.548	<b>0.460±0.142</b>	0.459-0.461
	Simple filling	0.687±0.094	0.686-0.688	0.761±0.076	0.760-0.761	<b>0.455±0.180</b>	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
	<i>p</i> value	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>	
Data sampling											
	No sampling	<b>0.738±0.101</b>	0.737-0.739	<b>0.823±0.050</b>	0.822-0.823	<b>0.585±0.229</b>	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	<b>0.457±0.135</b>	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	<b>0.596±0.161</b>	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
	<i>p</i> value	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>	

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	<b>0.713±0.105</b>	0.712-0.713	0.761±0.074	0.760-0.761	0.447±0.173	0.445-0.448	<b>0.513±0.177</b>	0.512-0.514	0.448±0.128	0.447-0.449
Boruta selection	0.706±0.103	0.705-0.707	<b>0.766±0.073</b>	0.765-0.766	<b>0.449±0.170</b>	0.448-0.450	0.501±0.166	0.500-0.503	<b>0.450±0.127</b>	0.449-0.451
<i>p</i> value	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>	

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

For peer review only

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

machine learning algorithms	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
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±0.142	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	<b>0.793±0.083</b>	0.791-0.794	<b>0.810±0.058</b>	0.809-0.811	<b>0.545±0.157</b>	0.543-0.548	<b>0.576±0.162</b>	0.573-0.579	<b>0.537±0.108</b>	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



5	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±0.075	0.567-0.573	0.467±0.111	0.465-0.469
7	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±0.056	0.555-0.560	0.421±0.107	0.419-0.423
9	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±0.088	0.433-0.440	0.411±0.152	0.408-0.413
11	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±0.065	0.427-0.432	0.444±0.119	0.441-0.446
13	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±0.013	0.503-0.511	0.434±0.141	0.432-0.437
15	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±0.065	0.440-0.446	0.449±0.115	0.447-0.451
17	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±0.065	0.510-0.516	0.486±0.112	0.484-0.488
19	<i>p</i> value		<b><i>p</i>&lt;0.0001</b>	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>	

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

207 **Table 3** Predictive performance indicators of the 5 best models

	AUC	accuracy	precision	recall rate	F1 value
model 1	<b>0.9141</b>	<b>0.8947</b>	<b>0.75</b>	0.6667	<b>0.7059</b>
model 2	0.9055	0.8105	0.5	<b>0.7778</b>	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

208 **DISCUSSION**

209 Traditional Chinese medicine has been used for the prevention and treatment of diseases  
 210 for centuries <sup>11</sup>. In recent years, the application of Chinese herbal containing Panax  
 211 notoginseng saponin, including injections, in clinical practice has become more and  
 212 more common, while the ADR often causes concerns. Studies have shown that the  
 213 Chinese herbal ingredients, traditional Chinese medicine preparation and combination  
 214 medication are the important factors for the ADR of Chinese herbal injections  
 215 containing Panax notoginseng saponin. Drug eruption (50.5%), allergic reactions  
 216 (20.4%) and anaphylactic shock (9.7%) are the most common, and some cases are even  
 217 life-threatening <sup>5</sup>. However, the ADR monitoring methods, including spontaneous  
 218 reporting systems, prescription event monitoring and centralized hospital monitoring  
 219 system, are reported after the event, and may even have data bias, underreporting or  
 220 repeated reporting. Therefore, the realization of ADR prediction has important  
 221 significance for prevent and avoid ADR of Chinese herbal injections containing Panax  
 222 notoginseng saponin in clinical practice.

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

1  
2  
3  
4 224 data sets, which were from data filling, data sampling and variable selection, were  
5  
6 225 combined with 18 machine learning algorithms to establish 1080 ADR prediction  
7  
8  
9 226 models. The AUC, accuracy, precision, recall rate and F1 value were used to evaluate  
10  
11 227 the predictive performance of the models. According to the parameters of the best  
12  
13  
14 228 model, an ADR prediction system for the Chinese herbal injections containing Panax  
15  
16  
17 229 notoginseng saponin was developed. This predictive system had high accuracy and  
18  
19  
20 230 precision, and had potential value for clinical application.

21  
22 231 In recent years, some ADR prediction models based on data mining <sup>6-9</sup>, machine  
23  
24 232 learning algorithms <sup>10, 12-15</sup>, and statistical methods <sup>16-18</sup>, have been developed.  
25  
26  
27 233 Tangiisuran et al. <sup>16</sup> combined univariate analysis and multivariate binary logistic  
28  
29  
30 234 regression for the identification of clinical risk factors to develop an ADR risk model.  
31  
32 235 The AUC of the model at internal and external validation stage was 0.74 and 0.73,  
33  
34  
35 236 respectively, the sensitivity was 80% and 84%, and the specificity was 55% and 43%  
36  
37  
38 237 <sup>16</sup>. Imai et al. <sup>10</sup> used artificial neural networks to predict the ADR risk and produced an  
39  
40  
41 238 AUC of 0.83. Compared with these models, the model established in our study had  
42  
43  
44 239 better predictive performance (accuracy was 0.8947, precision was 0.75, recall rate was  
45  
46  
47 240 0.6667 and AUC was 0.914). As missing data is common in the real-world health  
48  
49  
50 241 system, the methods of data filling used in our study may be advantageous for the deal  
51  
52  
53 242 with imbalanced data in clinical real-world research. More importantly, the model with  
54  
55  
56 243 optimal predictive performance selected from the 1080 models, was used to develop  
57  
58  
59 244 the ADR risk prediction system, which is potentially convenient for clinical practice  
60  
245 because of its' simple operation, fast calculation, and high accuracy.

1  
2  
3  
4 246 It is worth noting that Hammann et al.<sup>19</sup> established a decision tree model based  
5  
6 247 on the chemical, physical, and structural properties of compounds for the prediction of  
7  
8  
9 248 ADR occurrence and the model had high predictive accuracies (78.9–90.2%).  
10  
11 249 Unfortunately, the model ignored the effect of pathological and physiological  
12  
13  
14 250 conditions and the combination medication on ADR. More than 80 factors including  
15  
16  
17 251 the patient's pathophysiological characteristics, clinical laboratory results, and  
18  
19  
20 252 medication conditions, were performed by 3 variable selection methods in our study.  
21  
22 253 Meanwhile, we using the SHAP value to explain the contribution of the variables to the  
23  
24  
25 254 model.

26  
27 255 The importance of the variable indicates that whether the patients have dermatoses  
28  
29  
30 256 will significantly affect the models' predictive performance. Cutaneous ADR is one of  
31  
32  
33 257 the most common adverse reactions of *Panax notoginseng*, such as erythema  
34  
35 258 multiforme, urticaria, severe erythema multiforme and acute generalized  
36  
37  
38 259 exanthematous pustulosis<sup>20, 21</sup>. Therefore, those patients with original dermatoses are  
39  
40  
41 260 more likely to have ADR after using *Panax notoginseng*. In addition, we found that the  
42  
43  
44 261 age and gender are related to the occurrence of *Panax notoginseng*-induced ADR, which  
45  
46 262 is consistent with the results reported by Yang et al.<sup>22</sup>.

47  
48 263 However, our data were all from southwest China, and more data were needed to  
49  
50  
51 264 further evaluate the model prediction performance. In addition, a prospective controlled  
52  
53  
54 265 trial is required to demonstrate the accuracy of the ADR prediction system.

55  
56 266 **Contributors** XWW, EWL and RST were involved in the conception and design of  
57  
58 267 the study. XWW drafted the article. JYZ, HC, XWS and YLW analyzed the data.  
59  
60

1  
2  
3 268 EWL and RST revised the manuscript. All authors gave final approval of the version  
4  
5 269 to be published. The corresponding author attests that all listed authors meet  
6  
7 270 authorship criteria and that no others meeting the criteria have been omitted. RST is  
8  
9 271 the guarantor.  
10  
11  
12

13 272 **Funding** This study was funded by the National Natural Science Foundation of China  
14  
15 273 (No. 72004020), the Key Research and Development Program of Science and  
16  
17 274 Technology Department of Sichuan Province (No. 2019YFS0514), the Postgraduate  
18  
19 275 Research and Teaching Reform Project of the University of Electronic Science and  
20  
21 276 Technology of China (No. JYJG201919) and the Research Subject of Health  
22  
23 277 Commission of Sichuan Province (No. 19PJ262).  
24  
25  
26  
27

28 278 **Competing interests** None declared.  
29  
30

31 279 **Patient consent for publication** Not required.  
32  
33

34 280 **Ethics approval** Ethical approval: This study was approved by the Ethics Committee  
35  
36 281 of Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital  
37  
38 282 (2017-11-01).  
39  
40  
41

42 283 **Provenance and peer review** Not commissioned; externally peer reviewed.  
43  
44

45 284 **Data availability statement** Data are available upon reasonable request. Data may be  
46  
47 285 obtained from a third party and are not publicly available. The first author  
48  
49 286 (7190175@uestc.edu.cn) will share any publicly available data if requested by email.  
50  
51  
52

53 287 **Supplementary material** This content has been supplied by the author(s). It has not  
54  
55 288 been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-  
56  
57 289 reviewed. Any opinions or recommendations discussed are solely those of the  
58  
59  
60

1  
2  
3 290 author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility  
4  
5 291 arising from any reliance placed on the content. Where the content includes any  
6  
7 292 translated material, BMJ does not warrant the accuracy and reliability of the  
8  
9 293 translations (including but not limited to local regulations, clinical guidelines,  
10  
11 294 terminology, drug names and drug dosages), and is not responsible for any error  
12  
13 295 and/or omissions arising from translation and adaptation or otherwise.  
14  
15  
16  
17

18 296 **Open access** This is an open access article distributed in accordance with the Creative  
19  
20 297 Commons Attribution Non Commercial (CC BY-NC4.0) license, which permits  
21  
22 298 others to distribute, remix, adapt, build upon this work non-commercially, and license  
23  
24 299 their derivative works on different terms, provided the original work is properly cited,  
25  
26 300 appropriate credit is given, any changes made indicated, and the use is non-  
27  
28 301 commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.  
29  
30  
31

## 32 302 REFERENCES

- 33  
34  
35 303 1 Xie W, Meng X, Zhai Y ,et al. Panax Notoginseng Saponins: A Review of Its  
36  
37 304 Mechanisms of Antidepressant or Anxiolytic Effects and Network Analysis on  
38  
39 305 Phytochemistry and Pharmacology. *Molecules* 2018; 23.  
40  
41  
42 306 2 Kim JH. Pharmacological and medical applications of Panax ginseng and  
43  
44 307 ginsenosides: a review for use in cardiovascular diseases. *J Ginseng Res* 2018;  
45  
46 308 42:264-269.  
47  
48  
49 309 3 Yang F, Ma Q, Matsabisa MG ,et al. Panax notoginseng for Cerebral  
50  
51 310 Ischemia: A Systematic Review. *Am J Chin Med* 2020; 48:1331-1351.  
52  
53  
54 311 4 Qu J, Xu N, Zhang J ,et al. Panax notoginseng saponins and their applications  
55  
56 312 in nervous system disorders: a narrative review. *Ann Transl Med* 2020;  
57  
58  
59  
60

- 1  
2  
3  
4 313 8:1525.  
5  
6 314 5 Xiang Z, Qiao T, Xiao H ,et al. The anaphylactoid constituents in Xue-Sai-  
7  
8  
9 315 Tong injection. *Planta Med* 2013; 79:1043-1050.  
10  
11 316 6 Wei L, Scott J. Association rule mining in the US Vaccine Adverse Event  
12  
13  
14 317 Reporting System (VAERS). *Pharmacoepidemiol Drug Saf* 2015; 24:922-933.  
15  
16  
17 318 7 Harpaz R, DuMouchel W, Shah NH ,et al. Novel data-mining methodologies  
18  
19 319 for adverse drug event discovery and analysis. *Clin Pharmacol Ther* 2012;  
20  
21 320 91:1010-1021.  
22  
23  
24 321 8 Sakaeda T, Tamon A, Kadoyama K ,et al. Data mining of the public version of  
25  
26 322 the FDA Adverse Event Reporting System. *Int J Med Sci* 2013; 10:796-803.  
27  
28  
29 323 9 Kadoyama K, Kuwahara A, Yamamori M ,et al. Hypersensitivity reactions to  
30  
31 324 anticancer agents: data mining of the public version of the FDA adverse event  
32  
33 325 reporting system, AERS. *J Exp Clin Cancer Res* 2011; 30:93.  
34  
35  
36  
37 326 10 Imai S, Takekuma Y, Kashiwagi H ,et al. Validation of the usefulness of  
38  
39 327 artificial neural networks for risk prediction of adverse drug reactions used for  
40  
41 328 individual patients in clinical practice. *PLoS One* 2020; 15:e0236789.  
42  
43  
44 329 11 Liu SH, Chuang WC, Lam W ,et al. Safety surveillance of traditional Chinese  
45  
46 330 medicine: current and future. *Drug Saf* 2015; 38:117-128.  
47  
48  
49 331 12 Choudhury O, Park Y, Salonidis T ,et al. Predicting Adverse Drug Reactions  
50  
51 332 on Distributed Health Data using Federated Learning. *AMIA Annu Symp Proc*  
52  
53 333 2019; 2019:313-322.  
54  
55  
56  
57 334 13 Liu X, Chen H. A research framework for pharmacovigilance in health social  
58  
59  
60

- 1  
2  
3  
4 335 media: Identification and evaluation of patient adverse drug event reports. *J*  
5  
6 336 *Biomed Inform* 2015; 58:268-279.  
7  
8  
9 337 14 Davis J, Costa VS, Peissig P ,et al. Demand-Driven Clustering in Relational  
10  
11 338 Domains for Predicting Adverse Drug Events. *Proc Int Conf Mach Learn*  
12  
13 339 2012; 2012:1287-1294.  
14  
15  
16 340 15 Lee CY, Chen YP. Prediction of drug adverse events using deep learning in  
17  
18 341 pharmaceutical discovery. *Brief Bioinform* 2021; 22:1884-1901.  
19  
20  
21 342 16 Tangiisuran B, Scutt G, Stevenson J ,et al. Development and validation of a  
22  
23 343 risk model for predicting adverse drug reactions in older people during  
24  
25 344 hospital stay: Brighton Adverse Drug Reactions Risk (BADRI) model. *PLoS*  
26  
27 345 *One* 2014; 9:e111254.  
28  
29  
30 346 17 Clothier HJ, Lawrie J, Lewis G ,et al. SAEFVIC: Surveillance of adverse  
31  
32 347 events following immunisation (AEFI) in Victoria, Australia, 2018. *Commun*  
33  
34 348 *Dis Intell (2018)* 2020; 44.  
35  
36  
37 349 18 Alvarez Y, Hidalgo A, Maignen F ,et al. Validation of statistical signal  
38  
39 350 detection procedures in eudravigilance post-authorization data: a retrospective  
40  
41 351 evaluation of the potential for earlier signalling. *Drug Saf* 2010; 33:475-487.  
42  
43  
44 352 19 Hammann F, Gutmann H, Vogt N ,et al. Prediction of adverse drug reactions  
45  
46 353 using decision tree modeling. *Clin Pharmacol Ther* 2010; 88:52-59.  
47  
48  
49 354 20 Yan S, Xiong H, Shao F ,et al. HLA-C\*12:02 is strongly associated with  
50  
51 355 Xuesaitong-induced cutaneous adverse drug reactions. *Pharmacogenomics J*  
52  
53 356 2019; 19:277-285.  
54  
55  
56  
57  
58  
59  
60



- 1  
2  
3  
4 357 21 Chen WJ, Kuang YY, Li JT. Analysis on 13 Cases of Adverse Drug Reaction  
5  
6 358 by Xuesaitong Injection. *Journal of North Pharmacy* 2013; 10:16-17.  
7  
8  
9 359 22 Yang P, Qian N, Yao D ,et al. 62 Cases of Adverse Reactions in Xuesaitong  
10  
11 360 Oral Preparations. *Chinese Medicine Modern Distance Education of China*  
12  
13  
14 361 2021; 19:34-36.  
15  
16

362

17  
18  
19 363 **Figure 1** ROC curve of the 5 best models.

20  
21  
22 364 **Figure 2** Importance matrix plot of each variable to the final prediction model.

23  
24 365 Variable names were shown in Supplementary Table 1. X83, pre-treatment serum

25  
26 366 levels; X55, renal function; X25, dermatoses; X1, gender; X2, age; X29, dose; X62,

27  
28 367 low-density lipoprotein; X64, hypoproteinemia; X30, anti-infective agents; X82, pre-

29  
30 368 treatment indicators of carcinoma; X79, hemoglobin; X6, history of allergy; X16,

31  
32 369 respiratory diseases; X66, albumin/globulin; X78, red blood cell; X81, hypersensitive

33  
34 370 C-reactive protein; X51, dermatology medication; X77, eosinophils; X13, Charlson

35  
36 371 comorbidity index (Score); X57, serum potassium.

37  
38 372 **Figure 3** SHAP summary plot of the top 20 variables of the model. Red represents

39  
40 373 higher variable values, and blue represents lower variable values. Variable names

41  
42 374 were shown in Supplementary Table 1. X83, pre-treatment serum levels; X55, renal

43  
44 375 function; X25, dermatoses; X1, gender; X2, age; X29, dose; X62, low-density

45  
46 376 lipoprotein; X64, hypoproteinemia; X30, anti-infective agents; X82, pre-treatment

47  
48 377 indicators of carcinoma; X79, hemoglobin; X6, history of allergy; X16, respiratory

49  
50 378 diseases; X66, albumin/globulin; X78, red blood cell; X81, hypersensitive C-reactive

1  
2  
3  
4 379 protein; X51, dermatology medication; X77, eosinophils; X13, Charlson comorbidity

5  
6 380 index (Score); X57, serum potassium.

7  
8  
9 381 **Figure 4** Sample size validation. The vertical bars represent the 95% confidence

10  
11 382 interval (CI) of AUC of ROC.

12  
13  
14 383 **Figure 5** The development of ADR prediction system.

15  
16 384 **Figure 6** The operation (A) and output (B) of the ADR prediction system.

17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For peer review only

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46

385

For peer review only

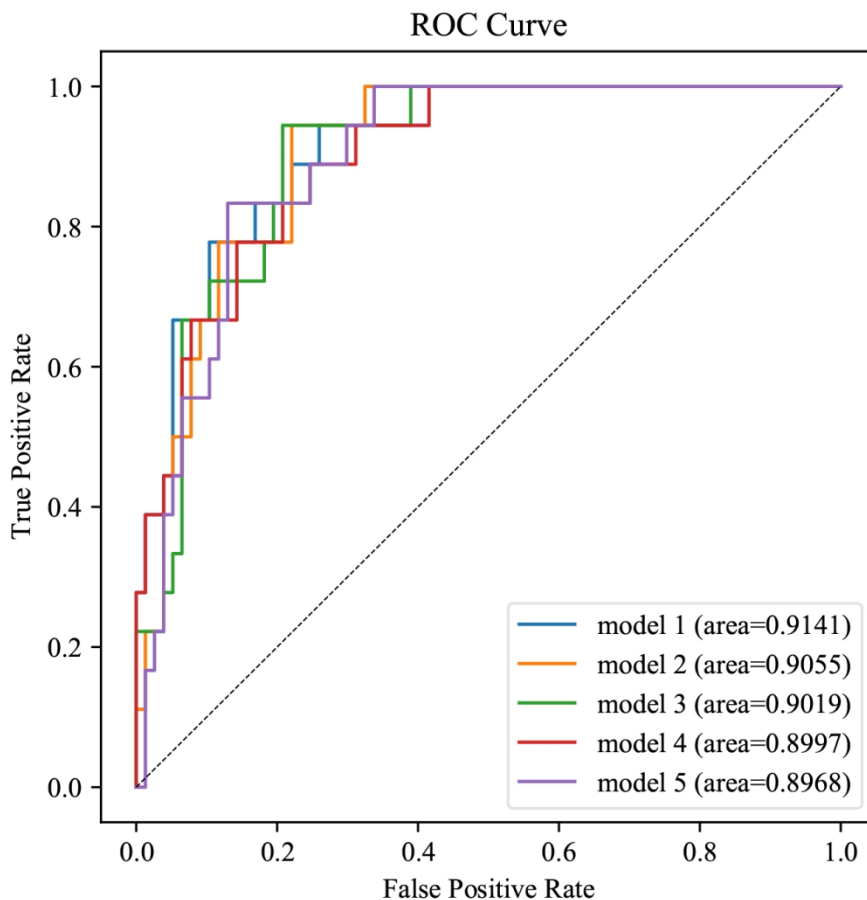


Figure 1 ROC curve of the 5 best models.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

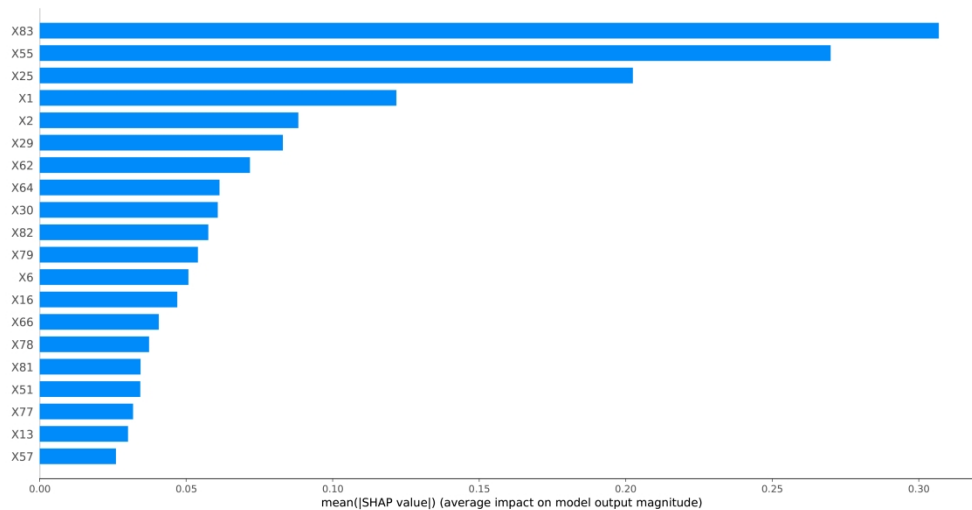


Figure 2 Importance matrix plot of each variable to the final prediction model.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

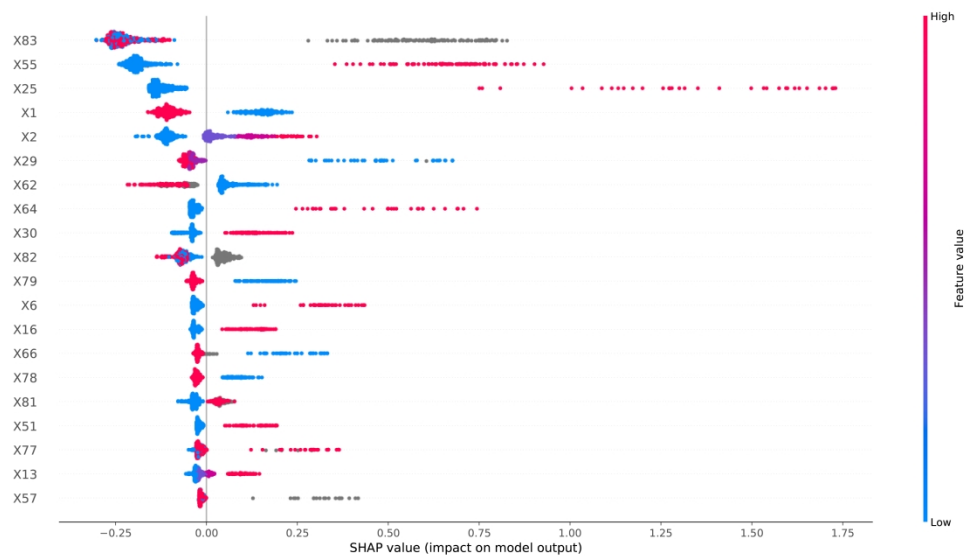


Figure 3 SHAP summary plot of the top 20 variables of the model.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

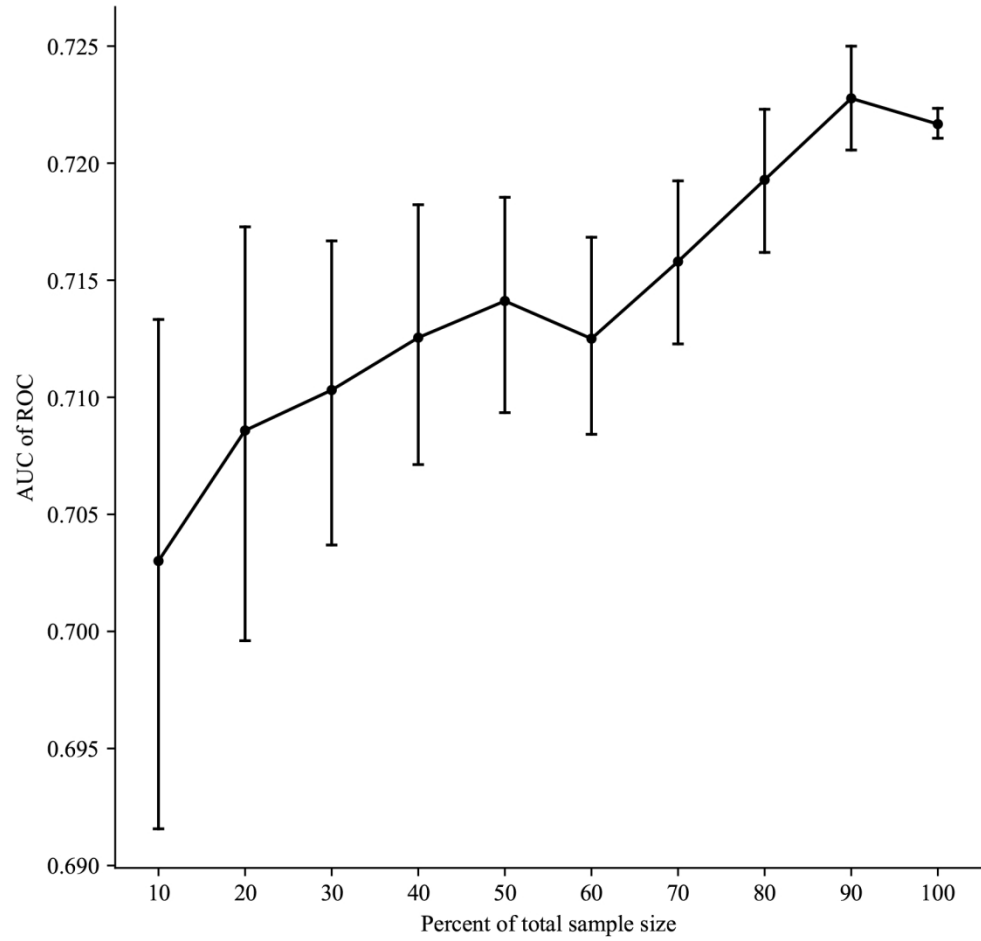


Figure 4 Sample size validation.

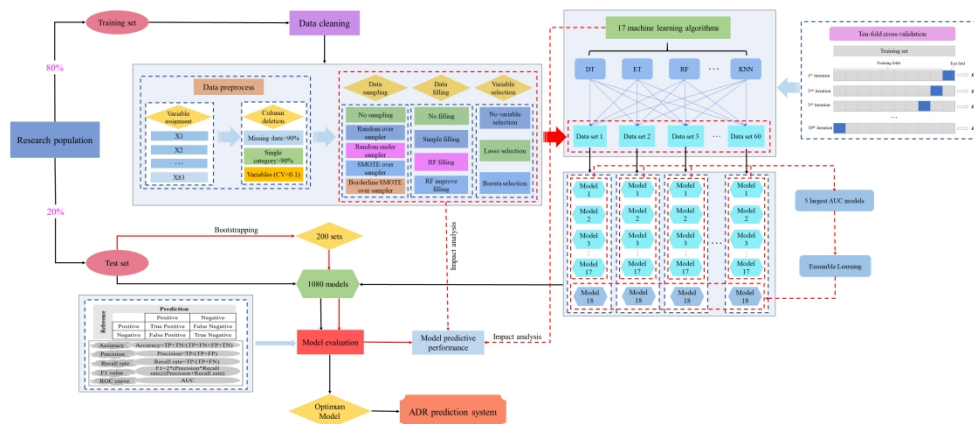


Figure 5 The development of ADR prediction system.

1050x472mm (96 x 96 DPI)



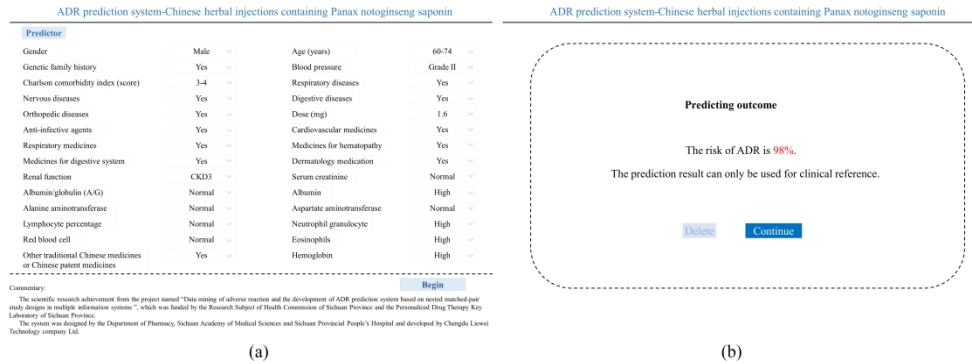


Figure 6 The operation (A) and output (B) of the ADR prediction system.

1899x688mm (96 x 96 DPI)

**Table 1** Variable assignment

Number	Variable	Assignment
	Adverse drug reaction	1, Yes; 0, No
X1	Gender	1, Male; 0, Female
X2	Age (years)	1, $\leq 44$ ; 2, $45 \leq \text{Age} \leq 59$ ; 3, $60 \leq \text{Age} \leq 74$ ; 4, $\geq 75$
X3	Body mass index (BMI, $\text{kg}/\text{m}^2$ )	1, $< 18.5$ ; 2, $18.5 \leq \text{BMI} \leq 23.9$ ; 3, $\geq 24$
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 ( $^{\circ}\text{C}$ )	1, $< 36.1$ ; 2, $36.1 \leq \text{Temperature} \leq 37.2$ ; 3, $> 37.3$
X10	Pulse (beats/min)	1, $< 60$ ; 2, $60 \leq \text{Pulse} \leq 100$ ; 3, $> 100$
X11	Breathe (times/min)	1, $< 12$ ; 2, $12 \leq \text{Breathe} \leq 20$ ; 3, $> 20$
X12	Blood pressure	0, Normal (systolic pressure $\leq 139$ mmHg or diastolic pressure $\leq 89$ mmHg); 1, Grade I ( $140$ mmHg $\leq$ systolic pressure $\leq 159$ mmHg or $90$ mmHg $\leq$ diastolic pressure $\leq 99$ mmHg); 2, Grade II ( $160$ mmHg $\leq$ systolic pressure $\leq 179$ mmHg or $100$ mmHg $\leq$ diastolic pressure $\leq 109$ mmHg); 3, Grade III (systolic pressure $\geq 180$ mmHg or diastolic pressure $\geq 110$ mmHg)
X13	Charlson comorbidity index (Score)	1, 0; 2, 1 or 2; 3, 3 or 4; 4, $\geq 5$
X14	Cardiovascular disease	1, Yes; 0, No

1			
2			
3			
4			
5	X15	Endocrine diseases	1, Yes; 0, No
6	X16	Respiratory diseases	1, Yes; 0, No
7	X17	Nervous diseases	1, Yes; 0, No
8	X18	Digestive diseases	1, Yes; 0, No
9	X19	Neoplastic diseases	1, Yes; 0, No
10	X20	Orthopedic diseases	1, Yes; 0, No
11	X21	Genito-urinary diseases	1, Yes; 0, No
12	X22	Hematopathy	1, Yes; 0, No
13	X23	Oculopathy	1, Yes; 0, No
14	X24	Ear-nose-throat diseases	1, Yes; 0, No
15	X25	Dermatoses	1, Yes; 0, No
16	X26	Immune rheumatism	1, Yes; 0, No
17	X27	Other diseases	1, Yes; 0, No
18	X28	Solvent	1, 0.9% sodium chloride injection; 2, 5% glucose injection; 3, Other solvents
19	X29	Dose (mg)	1, < 1.6; 2, =1.6; 3, > 1.6
20	X30	Anti-infective agents	1, Yes; 0, No
21	X31	Cardiovascular medicines	1, Yes; 0, No
22	X32	Medicines for digestive system	1, Yes; 0, No
23	X33	Respiratory medicines	1, Yes; 0, No
24	X34	Nervous system medicines	1, Yes; 0, No
25	X35	Medication in mental disorders	1, Yes; 0, No
26			
27			
28			
29			
30			
31			
32			
33			
34			
35			
36			
37			
38			
39			
40			
41			
42			
43			
44			
45			
46			

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

		or Chinese patent medicines	
1			
2			
3			
4			
5			
6			
7	X53	Urea	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
8	X54	Serum creatinine	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
9	X55	Renal function	1, Glomerular filtration rate $\geq 90$ ml/(min $\cdot$ 1.73m $^2$ ); 2, 60ml/(min $\cdot$ 1.73m $^2$ ) $\leq$ Glomerular filtration rate $\leq 89$ ml/(min $\cdot$ 1.73m $^2$ ); 3, 60ml/(min $\cdot$ 1.73m $^2$ ) $\leq$ Glomerular filtration rate $\leq 59$ ml/(min $\cdot$ 1.73m $^2$ ); 4, 5ml/(min $\cdot$ 1.73m $^2$ ) $\leq$ Glomerular filtration rate $\leq 29$ ml/(min $\cdot$ 1.73m $^2$ ); 5, Glomerular filtration rate $< 15$ ml/(min $\cdot$ 1.73m $^2$ )
11			
12			
13			
14			
15			
16			
17			
18	X56	Blood glucose	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
19	X57	Serum potassium	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
20	X58	Serum sodium	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
21	X59	Total cholesterol	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
22	X60	Triglyceride	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
23	X61	High-density lipoprotein	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
24	X62	Low-density lipoprotein	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
25	X63	Albumin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
26	X64	Hypoproteinemia	1, Yes; 0, No
27	X65	Globulin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
28	X66	Albumin/globulin (A/G)	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
29	X67	Aspartate aminotransferase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
30	X68	Alanine aminotransferase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
31			
32			
33			
34			
35			
36			
37			
38			
39			
40			
41			
42			
43			
44			
45			
46			

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

**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	<b>0.802±0.157</b>	0.796-0.808	0.749±0.237	0.740-0.759	<b>0.757±0.189</b>	0.750-0.764
RF improve filling	<b>0.887±0.094</b>	0.883-0.890	<b>0.832±0.096</b>	0.828-0.835	0.799±0.158	0.793-0.806	<b>0.751±0.240</b>	0.742-0.760	0.757±0.191	0.749-0.764
<i>p</i> value	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>	
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	<b>0.923±0.063</b>	0.920-0.925	0.858±0.085	0.854-0.861	<b>0.849±0.079</b>	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	<b>0.859±0.085</b>	0.855-0.862	0.841±0.074	0.837-0.844	<b>0.885±0.130</b>	0.879-0.890	<b>0.859±0.093</b>	0.855-0.863
<i>p</i> value	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>	
Variable selection										

BMJ Open: first published as 10.1136/bmjopen-2022-061457 on 08 September 2022. Downloaded from <http://bmjopen.bmj.com/> on October 30, 2024 by guest. Protected by copyright.

	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	<b>0.889±0.089</b>	0.886-0.892	<b>0.835±0.090</b>	0.832-0.838	<b>0.801±0.165</b>	0.796-0.807	<b>0.751±0.240</b>	0.743-0.759	<b>0.758±0.196</b>	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
	<i>p</i> value	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>	
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



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	<b>0.926±0.086</b>	0.919-0.933	<b>0.875±0.096</b>	0.867-0.883	<b>0.858±0.144</b>	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	<b>0.810±0.229</b>	0.792-0.828	<b>0.808±0.185</b>	0.793-0.822
<i>p</i> value		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>

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.



# 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 TRIPOD reporting 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
<b>Title</b>		
	<a href="#">#1</a> Identify the study as developing and / or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1

## Abstract

1		<a href="#">#2</a>	Provide a summary of objectives, study design, setting,	2
2				
3				
4			participants, sample size, predictors, outcome, statistical	
5				
6			analysis, results, and conclusions.	
7				
8				
9	<b>Introduction</b>			
10				
11				
12		<a href="#">#3a</a>	Explain the medical context (including whether diagnostic or	3
13				
14			prognostic) and rationale for developing or validating the	
15				
16			multivariable prediction model, including references to	
17				
18			existing models.	
19				
20				
21				
22		<a href="#">#3b</a>	Specify the objectives, including whether the study describes	4
23				
24			the development or validation of the model or both.	
25				
26				
27	<b>Methods</b>			
28				
29				
30	Source of data	<a href="#">#4a</a>	Describe the study design or source of data (e.g.,	5
31				
32			randomized trial, cohort, or registry data), separately for the	
33				
34			development and validation data sets, if applicable.	
35				
36				
37				
38	Source of data	<a href="#">#4b</a>	Specify the key study dates, including start of accrual; end of	5
39				
40			accrual; and, if applicable, end of follow-up.	
41				
42				
43	Participants	<a href="#">#5a</a>	Specify key elements of the study setting (e.g., primary care,	5
44				
45			secondary care, general population) including number and	
46				
47			location of centres.	
48				
49				
50				
51	Participants	<a href="#">#5b</a>	Describe eligibility criteria for participants.	5
52				
53				
54	Participants	<a href="#">#5c</a>	Give details of treatments received, if relevant	5
55				
56				
57				
58				
59				
60				

1	Outcome	<a href="#">#6a</a>	Clearly define the outcome that is predicted by the prediction	7
2			model, including how and when assessed.	
3				
4				
5				
6	Outcome	<a href="#">#6b</a>	Report any actions to blind assessment of the outcome to be	7
7			predicted.	
8				
9				
10				
11				
12	Predictors	<a href="#">#7a</a>	Clearly define all predictors used in developing or validating	6
13			the multivariable prediction model, including how and when	
14			they were measured	
15				
16				
17				
18				
19	Predictors	<a href="#">#7b</a>	Report any actions to blind assessment of predictors for the	6
20			outcome and other predictors.	
21				
22				
23				
24				
25	Sample size	<a href="#">#8</a>	Explain how the study size was arrived at.	5
26				
27				
28	Missing data	<a href="#">#9</a>	Describe how missing data were handled (e.g., complete-	6
29			case analysis, single imputation, multiple imputation) with	
30			details of any imputation method.	
31				
32				
33				
34				
35	Statistical	<a href="#">#10a</a>	If you are developing a prediction model describe how	6
36			predictors were handled in the analyses.	
37	analysis methods			
38				
39				
40				
41	Statistical	<a href="#">#10b</a>	If you are developing a prediction model, specify type of	7
42			model, all model-building procedures (including any	
43	analysis methods		predictor selection), and method for internal validation.	
44				
45				
46				
47				
48	Statistical	<a href="#">#10c</a>	If you are validating a prediction model, describe how the	7
49			predictions were calculated.	
50	analysis methods			
51				
52				
53				
54	Statistical	<a href="#">#10d</a>	Specify all measures used to assess model performance	7
55			and, if relevant, to compare multiple models.	
56	analysis methods			
57				
58				
59				
60				

1	Statistical	<a href="#">#10e</a>	If you are validating a prediction model, describe any model	7
2				
3	analysis methods		updating (e.g., recalibration) arising from the validation, if	
4			done	
5				
6				
7				
8				
9	Risk groups	<a href="#">#11</a>	Provide details on how risk groups were created, if done.	7
10				
11				
12	Development vs.	<a href="#">#12</a>	For validation, identify any differences from the development	7
13	validation		data in setting, eligibility criteria, outcome, and predictors.	
14				
15				
16				
17	<b>Results</b>			
18				
19				
20	Participants	<a href="#">#13a</a>	Describe the flow of participants through the study, including	8
21			the number of participants with and without the outcome	
22			and, if applicable, a summary of the follow-up time. A	
23			diagram may be helpful.	
24				
25				
26				
27				
28				
29				
30	Participants	<a href="#">#13b</a>	Describe the characteristics of the participants (basic	8
31			demographics, clinical features, available predictors),	
32			including the number of participants with missing data for	
33			predictors and outcome.	
34				
35				
36				
37				
38				
39				
40	Participants	<a href="#">#13c</a>	For validation, show a comparison with the development	8
41			data of the distribution of important variables (demographics,	
42			predictors and outcome).	
43				
44				
45				
46				
47				
48	Model	<a href="#">#14a</a>	If developing a model, specify the number of participants	9
49	development		and outcome events in each analysis.	
50				
51				
52				
53	Model	<a href="#">#14b</a>	If developing a model, report the unadjusted association, if	9
54	development		calculated between each candidate predictor and outcome.	
55				
56				
57				
58				
59				
60				

1	Model	<a href="#">#15a</a>	If developing a model, present the full prediction model to	9
2				
3	specification		allow predictions for individuals (i.e., all regression	
4			coefficients, and model intercept or baseline survival at a	
5			given time point).	
6				
7				
8				
9				
10				
11	Model	<a href="#">#15b</a>	If developing a prediction model, explain how to the use it.	9
12				
13	specification			
14				
15				
16	Model	<a href="#">#16</a>	Report performance measures (with CIs) for the prediction	9
17				
18	performance		model.	
19				
20				
21				
22	Model-updating	<a href="#">#17</a>	If validating a model, report the results from any model	9
23				
24			updating, if done (i.e., model specification, model	
25			performance).	
26				
27				
28				
29	<b>Discussion</b>			
30				
31				
32	Limitations	<a href="#">#18</a>	Discuss any limitations of the study (such as	17
33				
34			nonrepresentative sample, few events per predictor, missing	
35			data).	
36				
37				
38				
39				
40	Interpretation	<a href="#">#19a</a>	For validation, discuss the results with reference to	15
41				
42			performance in the development data, and any other	
43			validation data	
44				
45				
46				
47	Interpretation	<a href="#">#19b</a>	Give an overall interpretation of the results, considering	15
48				
49			objectives, limitations, results from similar studies, and other	
50			relevant evidence.	
51				
52				
53				
54				
55	Implications	<a href="#">#20</a>	Discuss the potential clinical use of the model and	16
56				
57			implications for future research	
58				
59				
60				

## Other information

- 1  
2  
3  
4 Supplementary [#21](#) Provide information about the availability of supplementary 18  
5  
6 information resources, such as study protocol, Web calculator, and data  
7  
8 sets.  
9  
10  
11  
12 Funding [#22](#) Give the source of funding and the role of the funders for the 18  
13  
14 present study.  
15  
16

17 None The TRIPOD checklist is distributed under the terms of the Creative Commons Attribution  
18 License CC-BY. This checklist can be completed online using <https://www.goodreports.org/>, a tool  
19  
20 made by the [EQUATOR Network](#) in collaboration with [Penelope.ai](#)  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



# BMJ Open

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

Journal:	<i>BMJ Open</i>
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
<b>Primary Subject Heading</b>:	Medical management
Secondary Subject Heading:	Medical management
Keywords:	Adverse events < THERAPEUTICS, Herbal medicine < THERAPEUTICS, Toxicity < THERAPEUTICS

SCHOLARONE™  
Manuscripts



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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12

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

13     Xing-Wei Wu<sup>1,2</sup>, Jia-Ying Zhang<sup>3</sup>, Huan Chang<sup>1</sup>, Xue-Wu Song<sup>1,2</sup>, Ya-Lin Wen<sup>1</sup>, En-  
14  
15     Wu Long<sup>1,2</sup>, Rong-Sheng Tong<sup>1,2</sup>

16  
17  
18     <sup>1</sup>Department of Pharmacy, Sichuan Provincial People's Hospital, University of  
19  
20  
21     Electronic Science and Technology of China, Chengdu, China,

22  
23     <sup>2</sup>Chinese Academy of Sciences Sichuan Translational Medicine Research  
24  
25  
26     Hospital, Chengdu 610072, China,

27  
28  
29     <sup>3</sup>Department of Pharmacy, Chengdu First People's Hospital, Chengdu 610095, China

30  
31  
32  
33

11    **Correspondence to**

34    Dr Rong-Sheng Tong, Department of Pharmacy, Sichuan Provincial People's Hospital,  
35  
36    University of Electronic Science and Technology of China, Chengdu, China. Chinese  
37  
38    Academy of Sciences Sichuan Translational Medicine Research Hospital, Chengdu  
39  
40  
41    610072, China. E-mail: [318004031@qq.com](mailto:318004031@qq.com)

42  
43  
44    **Word count:** 2349

45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4 17 **Develop an ADR prediction system of Chinese herbal injection**  
5  
6 18 **containing Panax notoginseng saponin: a nested case-control study**  
7  
8  
9 19 **using machine learning**  
10

11  
12 20 **ABSTRACT**  
13  
14

15 21 **Objective** This study aimed to develop an adverse drug reactions (ADR) antecedent  
16  
17  
18 22 prediction system using machine learning algorithms to provide the reference for  
19  
20  
21 23 security usage of Chinese herbal injections containing Panax notoginseng saponin in  
22  
23  
24 24 clinical practice.

25 25 **Design** A nested case-control study.

26 26 **Setting** National Center for ADR Monitoring and the Electronic Medical Record (EMR)  
27  
28  
29 27 system.

30 28 **Participants** All patients were from 5 medical institutions in Sichuan Province from  
31  
32  
33  
34 29 January 2010 to December 2018.

35  
36  
37  
38 30 **Main outcomes/measures** Data of patients with ADR who used Chinese herbal  
39  
40  
41 31 injections containing Panax notoginseng saponin was collected from the National  
42  
43  
44 32 Center for ADR Monitoring. A nested case-control study was used to randomly match  
45  
46  
47 33 patients without ADR from the EMR system by the ratio of 1:4. Eighteen machine  
48  
49  
50 34 learning algorithms were applied for the development of ADR prediction models. Area  
51  
52  
53 35 under curve (AUC), accuracy, precision, recall rate and F1 value were used to evaluate  
54  
55  
56 36 the predictive performance of the model. An ADR prediction system was established  
57  
58  
59 37 by the best model selected from the 1080 models.

60 38 **Results** A total of 530 patients from 5 medical institutions were included, and 1080

1  
2  
3  
4 39 ADR prediction models were developed. Among these models, the AUC of the best  
5  
6 40 capable one was 0.9141 and the accuracy was 0.8947. According to the best model, a  
7  
8  
9 41 prediction system, which can provide early identification of patients at risk for the ADR  
10  
11  
12 42 of Panax notoginseng saponin, has been established.

13  
14 43 **Conclusion** The prediction system developed based on the machine learning model in  
15  
16  
17 44 this study had good predictive performance and potential clinical application.

18  
19 45 **Key words** Adverse drug reactions, Chinese herbal injection, Machine learning,  
20  
21  
22 46 Prediction system, Panax notoginseng saponin

#### 23 24 25 47 **Strengths and limitations of this study**

- 26  
27 48 ➤ To the best of our knowledge, this study was the first to develop an ADR prediction  
28  
29  
30 49 system for Chinese herbal injection containing Panax notoginseng saponin using  
31  
32  
33 50 machine learning.
- 34  
35 51 ➤ Data of ADR patients came from the National Center for Adverse Drug Reaction  
36  
37  
38 52 Monitoring, which is highly representative.
- 39  
40 53 ➤ In order to obtain the best model, the data processing adopted 4 data filling, 5 data  
41  
42  
43 54 sampling, 3 variable selection methods, and 18 machine learning algorithms were  
44  
45  
46 55 applied for model establishment.
- 47  
48 56 ➤ The area under curve, accuracy, precision, recall rate, and F1 value were used to  
49  
50  
51 57 evaluate the predictive performance of the model.
- 52  
53 58 ➤ As the study population was all from southwest China, the results may be biased  
54  
55  
56 59 while the prediction system was applied in other medical institutions.  
57  
58  
59  
60

## 60 INTRODUCTION

61 Panax notoginseng saponins, as the main ingredients of Panax notoginseng (Buck.)  
62 F.H.Chen, has been widely used in the disease therapy of nervous system and cardio-  
63 cerebral vascular system <sup>1-4</sup>. High frequency of adverse drug reactions (ADR) in  
64 Chinese herbal containing Panax notoginseng saponin has received widespread  
65 attention. Among these ADR, about 69.57% were caused by injections, mainly  
66 manifested as drug eruption (50.5%), allergic reaction (20.4%) and anaphylactic shock  
67 (9.7%), which can be life-threatening in severe cases <sup>5</sup>.

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

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

1  
2  
3  
4 82 practice. In view of these challenges, this study aimed to develop an ADR prediction  
5  
6 83 system of Chinese herbal injections containing Panax notoginseng saponin based on  
7  
8  
9 84 machine learning algorithms and provide reference for clinical ADR management and  
10  
11  
12 85 prevention.

## 13 14 15 86 **METHODS**

### 16 17 18 87 **Data collection**

19  
20  
21 88 ADR patients who used Chinese herbal injections containing Panax notoginseng  
22  
23  
24 89 included in this study were from the National Center for Adverse Drug Reaction  
25  
26  
27 90 Monitoring reported by 5 hospitals in Sichuan Province from January 2010 to  
28  
29  
30 91 December 2018. Then, a nested case-control study was used to randomly match patients  
31  
32  
33 92 without ADR from the Electronic Medical Record (EMR) system of the 5 medical  
34  
35  
36 93 institutions. The ratio of patients with ADR to those without ADR was 1:4. For multiple  
37  
38  
39 94 lab results, in order to facilitate clinical application, we selected the last results of  
40  
41  
42 95 patients before the usage of medication. And for multiple admissions, all patients were  
43  
44  
45 96 included according to their first admission.

46  
47  
48 97 This study was approved by the Ethics Committee of Sichuan Academy of Medical  
49  
50  
51 98 Sciences and Sichuan Provincial People's Hospital. Due to the retrospective nature of  
52  
53  
54 99 the study, informed consent was waived. And we hid the patients' personal information  
55  
56  
57 100 during the study.

### 58 59 101 **Data cleaning**

60 102 *Variable assignment*

1  
2  
3  
4 103 Binary-state variables were directly assigned values of 0 or 1. According to whether in  
5  
6 104 the normal range, clinical laboratory variables were assigned values of 1, 2 and 3 (1,  
7  
8  
9 105 below the normal range; 2, within the normal range; and 3, above the normal range).

10 106 *Column deletion*

11  
12  
13  
14 107 Variables with missing data >90%, or a single category >90%, or the coefficient of  
15  
16  
17 108 variation (CV) <0.1 were deleted.

18  
19 109 *Data filling*

20  
21  
22 110 There are 4 ways to data filling. No filling: retained the original data. Simple filling:  
23  
24 111 missing data of continuous variables replaced by the mean or median, and categorical  
25  
26  
27 112 variables by the mode. Random Forest (RF) filling: used the RF model to predict and  
28  
29  
30 113 replace the missing data directly. RF improve filling: ordered variables based on the  
31  
32  
33 114 number of missing data that were replaced by RF filling next.

34  
35 115 *Data sampling*

36  
37  
38 116 No sampling: built models from the original data. Random over sampler: randomly  
39  
40  
41 117 replicated the data of fewer categories to match the sample size to that of more  
42  
43 118 categories. Random under sampler: deleted the data of more categories to match the  
44  
45  
46 119 sample size to that of fewer categories. Synthetic minority oversampling technique  
47  
48  
49 120 (SMOTE) over sampler: synthesize new data from a small amount of original data.  
50  
51 121 Borderline SMOTE over sampler: synthesize new data from borderline data.

52  
53 122 *Variable selection*

54  
55  
56 123 No variable selection or use Lasso or Boruta for variable selection.

57  
58  
59 124 **Model establishment**



1  
2  
3  
4 125 Through different data filling, data sampling and variable selection, 60 data sets were  
5  
6 126 obtained. Eighteen machine learning algorithms, including AdaBoost, Bagging,  
7  
8  
9 127 Bernoulli Naïve Bayes (Bernoulli NB), Decision Tree (DT), Extra Tree (ET), Gaussian  
10  
11 128 Naïve Bayes (Gaussian NB), Gradient Boosting, K-Nearest Neighbor (KNN), Latent  
12  
13  
14 129 Dirichlet Allocation (LDA), Logistic Regression (LR), Multinomial Naïve Bayes  
15  
16  
17 130 (Multinomial NB), Passive Aggressive, Quadratic Discriminant Analysis (QDA), RF,  
18  
19 131 Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), eXtreme  
20  
21  
22 132 Gradient Boosting (XGBoost), and Ensemble Learning, were used to build models.

23  
24  
25 133 The model establishment was as follows. The data were randomly divided into a  
26  
27 134 training set and a test set by the ratio of 8:2. The training set was used to build models,  
28  
29  
30 135 and the test set was used to evaluate the predictive performance of the models. Ten-fold  
31  
32 136 cross-validation on the training set was applied for internal validation of the model, and  
33  
34  
35 137 200 Bootstrapping samples from the test set for the evaluation of the impact of different  
36  
37 138 data processing methods or machine learning algorithms on model predictive  
38  
39  
40 139 performance. Ensemble learning models were developed by 5 machine learning  
41  
42  
43 140 algorithms with the largest area under curve (AUC) on each data set.

#### 44 45 46 141 **Model evaluation**

47  
48  
49 142 We used the AUC, accuracy, precision, recall rate, and F1 value to evaluate the  
50  
51 143 predictive performance of the model. Five models with the largest AUC were compared,  
52  
53  
54 144 and the best model was selected to develop an ADR prediction system of Chinese herbal  
55  
56  
57 145 injections containing Panax notoginseng saponin. SHapley Additive exPlanations  
58  
59 146 (SHAP) helped to explain the contribution of variables to the model.

### 147 **Sample size assessment**

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

### 153 **Patient and public involvement**

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

### 155 **Statistical Analysis**

156 Categorical variables were expressed as counts and percentages and continuous  
157 variables as mean  $\pm$  standard deviation. Analysis of variance will be used if the data  
158 were normally distributed and the variances were equal, otherwise, Kruskal-Wallis test  
159 will be used.  $p$  value  $< 0.05$  were considered statistically significant. Hypothesis testing  
160 and models building were implemented using the stats and sklearn packages in Python  
161 (Version 3.8), respectively.

## 162 **RESULTS**

### 163 **Research population**

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

### 167 **Data cleaning**

1  
2  
3  
4 168 The results of 83 variables assignment were shown in Supplementary Table 2. After the  
5  
6 169 column deletion, 63 variables were included in the following study (Supplementary  
7  
8  
9 170 Table 3). Then, 4 data filling methods were used for replacing the 1,290 (3.86%)  
10  
11 171 missing data. We used Lasso or Boruta for variable selection, and the results were  
12  
13  
14 172 shown in Supplementary Table 3. Using 4 data filling, 5 data sampling and 3 variable  
15  
16  
17 173 selection methods for data processing respectively, 60 data sets were obtained.

#### 174 **Model establishment**

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

#### 183 **Model evaluation**

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

#### 188 **Model interpretation**

1  
2  
3  
4 189 The importance of each variable to the final prediction model was shown in Figure 2.  
5  
6 190 The result showed that pre-treatment serum levels, renal function, dermatoses, gender  
7  
8  
9 191 and age were the top 5 most important variables for the model. We used the SHAP  
10  
11 192 value to explain the contribution of the variables to the model, and the SHAP value of  
12  
13  
14 193 the top 20 was shown in Figure 3. This plot explains how high and low variables values  
15  
16  
17 194 were in relation to SHAP values. For the prediction model, the higher the SHAP value  
18  
19 195 of a variable, the more likely ADR occurs.  
20  
21  
22

### 23 196 **Sample size assessment**

24  
25  
26 197 With the continuously increased size of sample data, the AUC values of the testing sets  
27  
28 198 continued to increase, which shows a sufficient sample size included in this study  
29  
30  
31 199 (Figure 4).  
32  
33

### 34 200 **Develop an ADR prediction system for Panax notoginseng saponin**

35  
36  
37 201 According to the best model, a prediction system for the ADR of Panax notoginseng  
38  
39 202 saponin has been developed and we had obtained the software copyright. The  
40  
41  
42 203 development of the ADR prediction system was shown in Figure 5. The operation and  
43  
44  
45 204 output of the system were shown in Figure 6.  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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	<b>0.786±0.101</b>	0.785-0.787	<b>0.770±0.070</b>	0.769-0.771	0.437±0.162	0.435-0.438	<b>0.546±0.208</b>	0.544-0.548	<b>0.460±0.142</b>	0.459-0.461
	Simple filling	0.687±0.094	0.686-0.688	0.761±0.076	0.760-0.761	<b>0.455±0.180</b>	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
	<i>p</i> value	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>	
Data sampling											
	No sampling	<b>0.738±0.101</b>	0.737-0.739	<b>0.823±0.050</b>	0.822-0.823	<b>0.585±0.229</b>	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	<b>0.457±0.135</b>	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	<b>0.596±0.161</b>	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
	<i>p</i> value	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>	

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46

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	<b>0.713±0.105</b>	0.712-0.713	0.761±0.074	0.760-0.761	0.447±0.173	0.445-0.448	<b>0.513±0.177</b>	0.512-0.514	0.448±0.128	0.447-0.449
Boruta selection	0.706±0.103	0.705-0.707	<b>0.766±0.073</b>	0.765-0.766	<b>0.449±0.170</b>	0.448-0.450	0.501±0.166	0.500-0.503	<b>0.450±0.127</b>	0.449-0.451
<i>p</i> value	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>	

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

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

machine learning algorithms	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
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±0.142	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	<b>0.793±0.083</b>	0.791-0.794	<b>0.810±0.058</b>	0.809-0.811	<b>0.545±0.157</b>	0.543-0.548	<b>0.576±0.162</b>	0.573-0.579	<b>0.537±0.108</b>	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

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±0.075	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±0.056	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±0.088	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±0.065	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±0.013	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±0.065	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±0.065	0.510-0.516	0.486±0.112	0.484-0.488
<i>p</i> value	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>	

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



211 **Table 3** Predictive performance indicators of the 5 best models

	AUC	accuracy	precision	recall rate	F1 value
model 1	<b>0.9141</b>	<b>0.8947</b>	<b>0.75</b>	0.6667	<b>0.7059</b>
model 2	0.9055	0.8105	0.5	<b>0.7778</b>	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

212 **DISCUSSION**

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

1  
2  
3  
4 227 In our study, a nested case-control study was performed for data collection. In  
5  
6 228 order to obtain the best model, we used 4 data filling, 5 data sampling and 3 variable  
7  
8  
9 229 selection methods for data processing, and combined 18 machine learning algorithms  
10  
11  
12 230 to establish 1080 ADR prediction models. By comparing the AUC, accuracy, precision,  
13  
14  
15 231 recall rate and F1 value of these models, the best one was selected to develop an ADR  
16  
17 232 prediction system for the Chinese herbal injections containing Panax notoginseng  
18  
19  
20 233 saponin.  
21  
22

23 234 In recent years, some ADR prediction models have been developed based on data  
24  
25 235 mining <sup>6-9</sup>, machine learning algorithms <sup>10, 12-15</sup>, and statistical methods <sup>16-18</sup>.  
26  
27  
28 236 Tangiisuran et al. <sup>16</sup> combined univariate analysis and multivariate binary logistic  
29  
30  
31 237 regression for the identification of clinical risk factors to develop an ADR risk model.  
32  
33  
34 238 The AUC of the model at the internal and external validation stage was 0.74 and 0.73,  
35  
36 239 respectively, the sensitivity was 80% and 84%, and the specificity was 55% and 43%  
37  
38  
39 240 <sup>16</sup>. Imai et al. <sup>10</sup> used artificial neural networks to predict the ADR risk and made an  
40  
41  
42 241 AUC of 0.83. Compared with other studies, the model established in our study had  
43  
44  
45 242 better predictive performance (accuracy was 0.8947, precision was 0.75, the recall rate  
46  
47 243 was 0.6667 and AUC was 0.914). As missing data is common in clinical practice, the  
48  
49  
50 244 methods of data filling used in our study may be advantageous for the deal with  
51  
52  
53 245 imbalanced data in clinical real-world research. More importantly, the system  
54  
55  
56 246 developed by the best model was potentially convenient for clinical application because  
57  
58  
59 247 of its' simple operation, fast calculation, and high accuracy.  
60

1  
2  
3  
4 248 It is worth noting that Hammann et al.<sup>19</sup> established a decision tree model based  
5  
6  
7 249 on the chemical, physical, and structural properties of compounds for the prediction of  
8  
9  
10 250 ADR occurrence and the model had high predictive accuracy (78.9–90.2%). However,  
11  
12 251 the model was difficult to interpret as it ignored the effect of pathological and  
13  
14  
15 252 physiological conditions and the combination medication on ADR. This made the  
16  
17  
18 253 model unlikely to be accepted by clinicians. In our study, we collected more than 80  
19  
20  
21 254 factors including the patient's pathophysiological characteristics, clinical laboratory  
22  
23  
24 255 results, and medication conditions. Meanwhile, the critical predictors associated with  
25  
26  
27 256 the ADR were identified by the SHAP values. Although using the SHAP values as a  
28  
29  
30 257 generalized approach to identify the important clinical determinants of ADR caused by  
31  
32  
33 258 Chinese herbal injections containing Panax notoginseng saponin is not possible, it may  
34  
35  
36 259 help generate clinical hypotheses for some specific clinical events.

37 260 The results of SHAP indicated that whether the patients have dermatoses will  
38  
39  
40 261 significantly affect the models' predictive performance. Cutaneous ADR is one of the  
41  
42  
43 262 most common adverse reactions of Panax notoginseng, such as erythema multiforme,  
44  
45  
46 263 urticaria, severe erythema multiforme and acute generalized exanthematous pustulosis  
47  
48  
49 264 <sup>20, 21</sup>. Therefore, those patients with original dermatoses are more likely to have ADR  
50  
51  
52 265 after using Panax notoginseng. In addition, we found that age and gender are related to  
53  
54  
55 266 the occurrence of Panax notoginseng-induced ADR, which is consistent with the results  
56  
57  
58 267 reported by Yang et al.<sup>22</sup>.

59 268 This study had some limitations. First, the small sample size of this study might  
60

1  
2  
3  
4 269 affect the model prediction performance. Second, as the study population was all from  
5  
6  
7 270 southwest China, the results may be biased while the prediction system was applied in  
8  
9  
10 271 other medical institutions. Finally, a prospective controlled trial is required to  
11  
12 272 demonstrate the accuracy of the ADR prediction system.

13  
14  
15 273 **Contributors** XWW, EWL and RST were involved in the conception and design of  
16  
17 274 the study. XWW drafted the article. JYZ, HC, XWS and YLW analyzed the data.  
18  
19  
20 275 EWL and RST revised the manuscript. All authors gave final approval of the version  
21  
22 276 to be published. The corresponding author attests that all listed authors meet  
23  
24 277 authorship criteria and that no others meeting the criteria have been omitted. RST is  
25  
26 278 the guarantor.

27  
28  
29  
30 279 **Funding** This study was funded by the National Natural Science Foundation of China  
31  
32 280 (No. 72004020), the Key Research and Development Program of Science and  
33  
34 281 Technology Department of Sichuan Province (No. 2019YFS0514), the Postgraduate  
35  
36 282 Research and Teaching Reform Project of the University of Electronic Science and  
37  
38 283 Technology of China (No. JYJG201919) and the Research Subject of Health  
39  
40 284 Commission of Sichuan Province (No. 19PJ262).

41  
42  
43  
44 285 **Competing interests** None declared.

45  
46  
47  
48 286 **Patient consent for publication** Not required.

49  
50  
51 287 **Ethics approval** Ethical approval: This study was approved by the Ethics Committee  
52  
53 288 of Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital  
54  
55 289 (2017-11-01).

56  
57  
58  
59 290 **Provenance and peer review** Not commissioned; externally peer reviewed.  
60

1  
2  
3 291 **Data availability statement** Data are available upon reasonable request. Data may be  
4  
5 292 obtained from a third party and are not publicly available. The first author  
6  
7 293 (7190175@uestc.edu.cn) will share any publicly available data if requested by email.  
8  
9

10  
11 294 **Supplementary material** This content has been supplied by the author(s). It has not  
12  
13 295 been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-  
14  
15 296 reviewed. Any opinions or recommendations discussed are solely those of the  
16  
17 297 author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility  
18  
19 298 arising from any reliance placed on the content. Where the content includes any  
20  
21 299 translated material, BMJ does not warrant the accuracy and reliability of the  
22  
23 300 translations (including but not limited to local regulations, clinical guidelines,  
24  
25 301 terminology, drug names and drug dosages), and is not responsible for any error  
26  
27 302 and/or omissions arising from translation and adaptation or otherwise.  
28  
29  
30  
31

32  
33 303 **Open access** This is an open access article distributed in accordance with the Creative  
34  
35 304 Commons Attribution Non Commercial (CC BY-NC4.0) license, which permits  
36  
37 305 others to distribute, remix, adapt, build upon this work non-commercially, and license  
38  
39 306 their derivative works on different terms, provided the original work is properly cited,  
40  
41 307 appropriate credit is given, any changes made indicated, and the use is non-  
42  
43 308 commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.  
44  
45  
46

## 47 309 **REFERENCES**

- 48  
49  
50 310 1 Xie W, Meng X, Zhai Y ,et al. Panax Notoginseng Saponins: A Review of Its  
51  
52 311 Mechanisms of Antidepressant or Anxiolytic Effects and Network Analysis on  
53  
54 312 Phytochemistry and Pharmacology. *Molecules* 2018; 23.  
55  
56  
57 313 2 Kim JH. Pharmacological and medical applications of Panax ginseng and  
58  
59  
60

- 1  
2  
3  
4 314 ginsenosides: a review for use in cardiovascular diseases. *J Ginseng Res* 2018;  
5  
6  
7 315 42:264-269.  
8  
9 316 3 Yang F, Ma Q, Matsabisa MG ,et al. Panax notoginseng for Cerebral  
10  
11  
12 317 Ischemia: A Systematic Review. *Am J Chin Med* 2020; 48:1331-1351.  
13  
14  
15 318 4 Qu J, Xu N, Zhang J ,et al. Panax notoginseng saponins and their applications  
16  
17  
18 319 in nervous system disorders: a narrative review. *Ann Transl Med* 2020;  
19  
20 320 8:1525.  
21  
22  
23 321 5 Xiang Z, Qiao T, Xiao H ,et al. The anaphylactoid constituents in Xue-Sai-  
24  
25  
26 322 Tong injection. *Planta Med* 2013; 79:1043-1050.  
27  
28  
29 323 6 Wei L, Scott J. Association rule mining in the US Vaccine Adverse Event  
30  
31 324 Reporting System (VAERS). *Pharmacoepidemiol Drug Saf* 2015; 24:922-933.  
32  
33  
34 325 7 Harpaz R, DuMouchel W, Shah NH ,et al. Novel data-mining methodologies  
35  
36  
37 326 for adverse drug event discovery and analysis. *Clin Pharmacol Ther* 2012;  
38  
39 327 91:1010-1021.  
40  
41  
42 328 8 Sakaeda T, Tamon A, Kadoyama K ,et al. Data mining of the public version of  
43  
44  
45 329 the FDA Adverse Event Reporting System. *Int J Med Sci* 2013; 10:796-803.  
46  
47  
48 330 9 Kadoyama K, Kuwahara A, Yamamori M ,et al. Hypersensitivity reactions to  
49  
50 331 anticancer agents: data mining of the public version of the FDA adverse event  
51  
52 332 reporting system, AERS. *J Exp Clin Cancer Res* 2011; 30:93.  
53  
54  
55 333 10 Imai S, Takekuma Y, Kashiwagi H ,et al. Validation of the usefulness of  
56  
57  
58 334 artificial neural networks for risk prediction of adverse drug reactions used for  
59  
60

- 1  
2  
3  
4 335 individual patients in clinical practice. *PLoS One* 2020; 15:e0236789.  
5  
6  
7 336 11 Liu SH, Chuang WC, Lam W ,et al. Safety surveillance of traditional Chinese  
8  
9 337 medicine: current and future. *Drug Saf* 2015; 38:117-128.  
10  
11  
12 338 12 Choudhury O, Park Y, Salonidis T ,et al. Predicting Adverse Drug Reactions  
13  
14 339 on Distributed Health Data using Federated Learning. *AMIA Annu Symp Proc*  
15  
16 340 2019; 2019:313-322.  
17  
18  
19  
20 341 13 Liu X, Chen H. A research framework for pharmacovigilance in health social  
21  
22 342 media: Identification and evaluation of patient adverse drug event reports. *J*  
23  
24 343 *Biomed Inform* 2015; 58:268-279.  
25  
26  
27  
28 344 14 Davis J, Costa VS, Peissig P ,et al. Demand-Driven Clustering in Relational  
29  
30 345 Domains for Predicting Adverse Drug Events. *Proc Int Conf Mach Learn*  
31  
32 346 2012; 2012:1287-1294.  
33  
34  
35  
36 347 15 Lee CY, Chen YP. Prediction of drug adverse events using deep learning in  
37  
38 348 pharmaceutical discovery. *Brief Bioinform* 2021; 22:1884-1901.  
39  
40  
41  
42 349 16 Tangiisuran B, Scutt G, Stevenson J ,et al. Development and validation of a  
43  
44 350 risk model for predicting adverse drug reactions in older people during  
45  
46 351 hospital stay: Brighton Adverse Drug Reactions Risk (BADRI) model. *PLoS*  
47  
48 352 *One* 2014; 9:e111254.  
49  
50  
51  
52 353 17 Clothier HJ, Lawrie J, Lewis G ,et al. SAEFVIC: Surveillance of adverse  
53  
54 354 events following immunisation (AEFI) in Victoria, Australia, 2018. *Commun*  
55  
56 355 *Dis Intell (2018)* 2020; 44.  
57  
58  
59  
60

- 1  
2  
3  
4 356 18 Alvarez Y, Hidalgo A, Maignen F ,et al. Validation of statistical signal  
5  
6 357 detection procedures in eudravigilance post-authorization data: a retrospective  
7  
8  
9 358 evaluation of the potential for earlier signalling. *Drug Saf* 2010; 33:475-487.  
10  
11  
12 359 19 Hammann F, Gutmann H, Vogt N ,et al. Prediction of adverse drug reactions  
13  
14 360 using decision tree modeling. *Clin Pharmacol Ther* 2010; 88:52-59.  
15  
16  
17 361 20 Yan S, Xiong H, Shao F ,et al. HLA-C\*12:02 is strongly associated with  
18  
19 362 Xuesaitong-induced cutaneous adverse drug reactions. *Pharmacogenomics J*  
20  
21 363 2019; 19:277-285.  
22  
23  
24 364 21 Chen WJ, Kuang YY, Li JT. Analysis on 13 Cases of Adverse Drug Reaction  
25  
26 365 by Xuesaitong Injection. *Journal of North Pharmacy* 2013; 10:16-17.  
27  
28  
29 366 22 Yang P, Qian N, Yao D ,et al. 62 Cases of Adverse Reactions in Xuesaitong  
30  
31 367 Oral Preparations. *Chinese Medicine Modern Distance Education of China*  
32  
33 368 2021; 19:34-36.  
34  
35  
36  
37  
38  
39  
40  
41

42 370 **Figure 1** ROC curve of the 5 best models.

43  
44 371 **Figure 2** Importance matrix plot of each variable to the final prediction model.

45  
46 372 Variable names were shown in Supplementary Table 2. X83, pre-treatment serum  
47  
48 373 levels; X55, renal function; X25, dermatoses; X1, gender; X2, age; X29, dose; X62,  
49  
50 374 low-density lipoprotein; X64, hypoproteinemia; X30, anti-infective agents; X82, pre-  
51  
52 375 treatment indicators of carcinoma; X79, hemoglobin; X6, history of allergy; X16,  
53  
54 376 respiratory diseases; X66, albumin/globulin; X78, red blood cell; X81, hypersensitive  
55  
56  
57  
58  
59  
60



1  
2  
3  
4 377 C-reactive protein; X51, dermatology medication; X77, eosinophils; X13, Charlson  
5  
6  
7 378 comorbidity index (Score); X57, serum potassium.

8  
9 379 **Figure 3** SHAP summary plot of the top 20 variables of the model. Red represents  
10  
11  
12 380 higher variable values, and blue represents lower variable values. Variable names  
13  
14  
15 381 were shown in Supplementary Table 2. X83, pre-treatment serum levels; X55, renal  
16  
17  
18 382 function; X25, dermatoses; X1, gender; X2, age; X29, dose; X62, low-density  
19  
20  
21 383 lipoprotein; X64, hypoproteinemia; X30, anti-infective agents; X82, pre-treatment  
22  
23  
24 384 indicators of carcinoma; X79, hemoglobin; X6, history of allergy; X16, respiratory  
25  
26  
27 385 diseases; X66, albumin/globulin; X78, red blood cell; X81, hypersensitive C-reactive  
28  
29  
30 386 protein; X51, dermatology medication; X77, eosinophils; X13, Charlson comorbidity  
31  
32  
33 387 index (Score); X57, serum potassium.

34 388 **Figure 4** Sample size validation. The vertical bars represent the 95% confidence  
35  
36  
37 389 interval (CI) of AUC of ROC.

38  
39 390 **Figure 5** The development of ADR prediction system.

40  
41  
42 391 **Figure 6** The operation (A) and output (B) of the ADR prediction system.  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

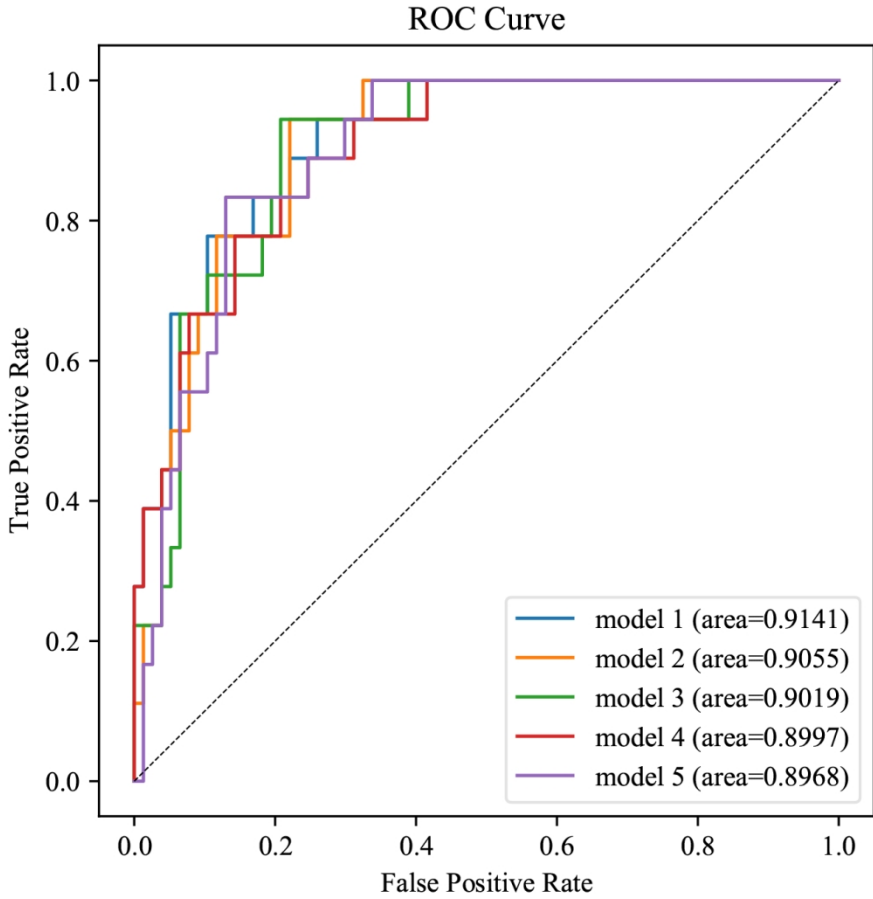


Figure 1 ROC curve of the 5 best models.

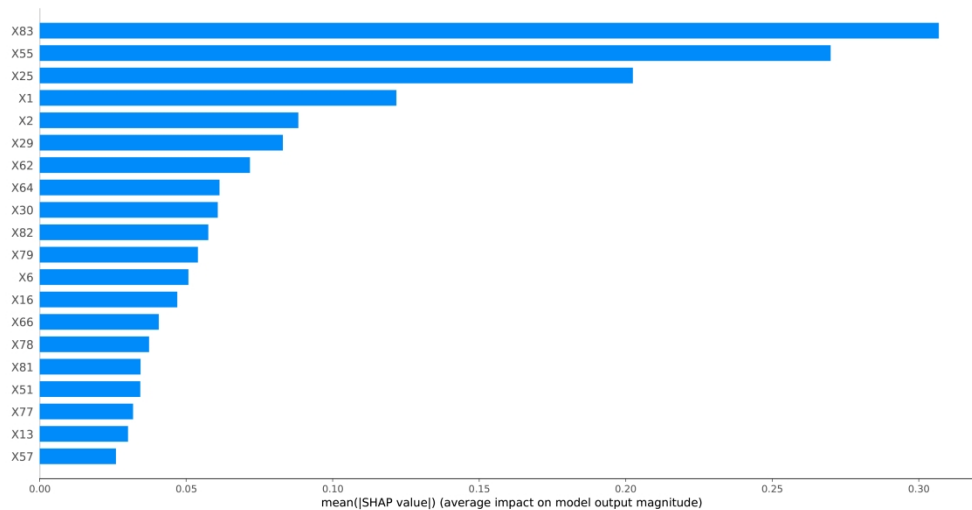


Figure 2 Importance matrix plot of each variable to the final prediction model.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

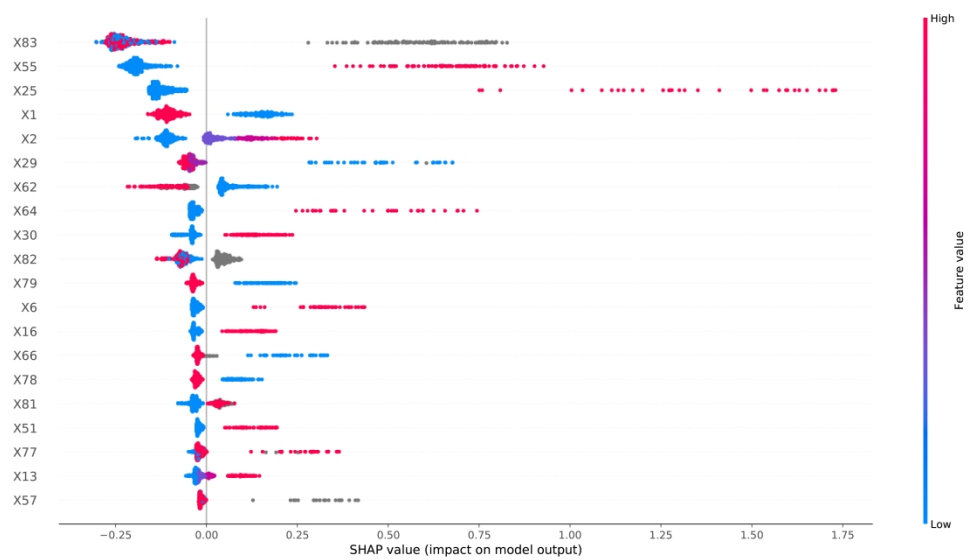


Figure 3 SHAP summary plot of the top 20 variables of the model.

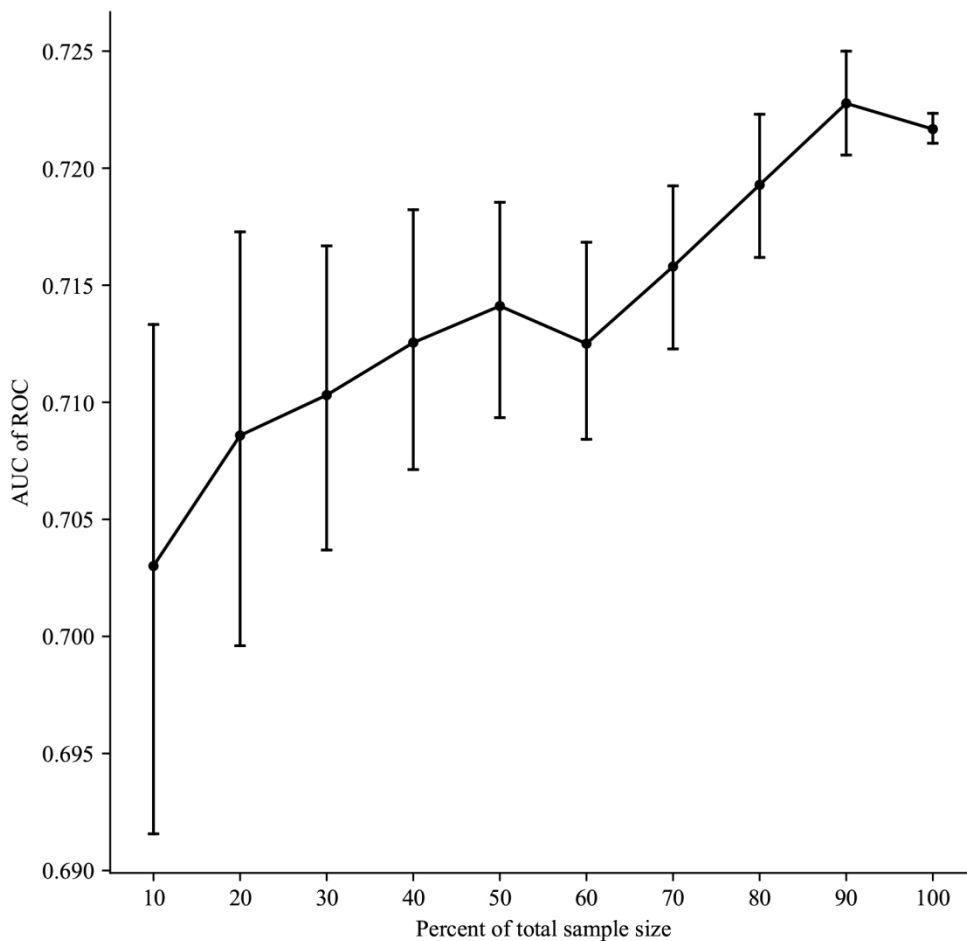


Figure 4 Sample size validation.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

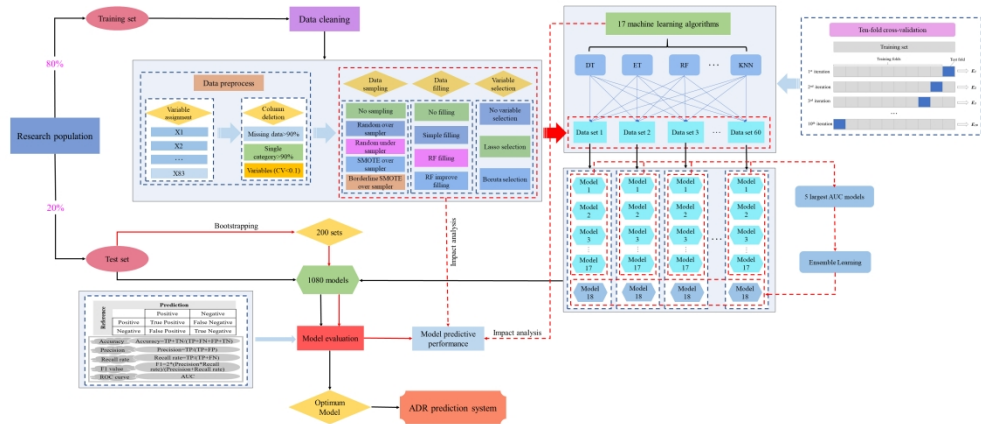


Figure 5 The development of ADR prediction system.

1050x472mm (96 x 96 DPI)

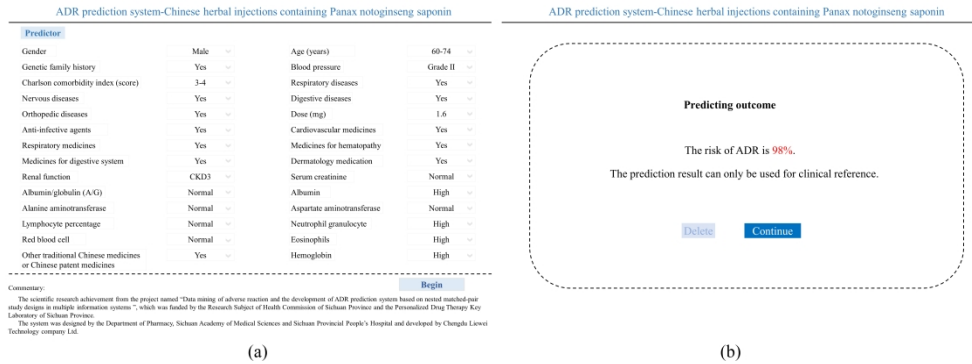


Figure 6 The operation (A) and output (B) of the ADR prediction system.

1899x688mm (96 x 96 DPI)

**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 <sup>2</sup> )	
< 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 (%)	



**Table 2** Variable assignment

Number	Variable	Assignment
	Adverse drug reaction	1, Yes; 0, No
X1	Gender	1, Male; 0, Female
X2	Age (years)	1, $\leq 44$ ; 2, $45 \leq \text{Age} \leq 59$ ; 3, $60 \leq \text{Age} \leq 74$ ; 4, $\geq 75$
X3	Body mass index (BMI, $\text{kg}/\text{m}^2$ )	1, $< 18.5$ ; 2, $18.5 \leq \text{BMI} \leq 23.9$ ; 3, $\geq 24$
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 ( $^{\circ}\text{C}$ )	1, $< 36.1$ ; 2, $36.1 \leq \text{Temperature} \leq 37.2$ ; 3, $> 37.3$
X10	Pulse (beats/min)	1, $< 60$ ; 2, $60 \leq \text{Pulse} \leq 100$ ; 3, $> 100$
X11	Breathe (times/min)	1, $< 12$ ; 2, $12 \leq \text{Breathe} \leq 20$ ; 3, $> 20$
X12	Blood pressure	0, Normal (systolic pressure $\leq 139$ mmHg or diastolic pressure $\leq 89$ mmHg); 1, Grade I ( $140$ mmHg $\leq$ systolic pressure $\leq 159$ mmHg or $90$ mmHg $\leq$ diastolic pressure $\leq 99$ mmHg); 2, Grade II ( $160$ mmHg $\leq$ systolic pressure $\leq 179$ mmHg or $100$ mmHg $\leq$ diastolic pressure $\leq 109$ mmHg); 3, Grade III (systolic pressure $\geq 180$ mmHg or diastolic pressure $\geq 110$ mmHg)
X13	Charlson comorbidity index (Score)	1, 0; 2, 1 or 2; 3, 3 or 4; 4, $\geq 5$
X14	Cardiovascular disease	1, Yes; 0, No

1			
2			
3			
4			
5	X15	Endocrine diseases	1, Yes; 0, No
6	X16	Respiratory diseases	1, Yes; 0, No
7	X17	Nervous diseases	1, Yes; 0, No
8	X18	Digestive diseases	1, Yes; 0, No
9	X19	Neoplastic diseases	1, Yes; 0, No
10	X20	Orthopedic diseases	1, Yes; 0, No
11	X21	Genito-urinary diseases	1, Yes; 0, No
12	X22	Hematopathy	1, Yes; 0, No
13	X23	Oculopathy	1, Yes; 0, No
14	X24	Ear-nose-throat diseases	1, Yes; 0, No
15	X25	Dermatoses	1, Yes; 0, No
16	X26	Immune rheumatism	1, Yes; 0, No
17	X27	Other diseases	1, Yes; 0, No
18	X28	Solvent	1, 0.9% sodium chloride injection; 2, 5% glucose injection; 3, Other solvents
19	X29	Dose (mg)	1, < 1.6; 2, =1.6; 3, > 1.6
20	X30	Anti-infective agents	1, Yes; 0, No
21	X31	Cardiovascular medicines	1, Yes; 0, No
22	X32	Medicines for digestive system	1, Yes; 0, No
23	X33	Respiratory medicines	1, Yes; 0, No
24	X34	Nervous system medicines	1, Yes; 0, No
25	X35	Medication in mental disorders	1, Yes; 0, No
26			
27			
28			
29			
30			
31			
32			
33			
34			
35			
36			
37			
38			
39			
40			
41			
42			
43			
44			
45			
46			

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

		or Chinese patent medicines	
1			
2			
3			
4			
5			
6			
7	X53	Urea	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
8	X54	Serum creatinine	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
9	X55	Renal function	1, Glomerular filtration rate $\geq 90$ ml/(min $\cdot$ 1.73m $^2$ ); 2, 60ml/(min $\cdot$ 1.73m $^2$ ) $\leq$ Glomerular filtration rate $\leq 89$ ml/(min $\cdot$ 1.73m $^2$ ); 3, 60ml/(min $\cdot$ 1.73m $^2$ ) $\leq$ Glomerular filtration rate $\leq 59$ ml/(min $\cdot$ 1.73m $^2$ ); 4, 5ml/(min $\cdot$ 1.73m $^2$ ) $\leq$ Glomerular filtration rate $\leq 29$ ml/(min $\cdot$ 1.73m $^2$ ); 5, Glomerular filtration rate $< 15$ ml/(min $\cdot$ 1.73m $^2$ )
11			
12			
13			
14			
15			
16			
17			
18	X56	Blood glucose	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
19	X57	Serum potassium	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
20	X58	Serum sodium	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
21	X59	Total cholesterol	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
22	X60	Triglyceride	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
23	X61	High-density lipoprotein	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
24	X62	Low-density lipoprotein	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
25	X63	Albumin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
26	X64	Hypoproteinemia	1, Yes; 0, No
27	X65	Globulin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
28	X66	Albumin/globulin (A/G)	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
29	X67	Aspartate aminotransferase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
30	X68	Alanine aminotransferase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
31			
32			
33			
34			
35			
36			
37			
38			
39			
40			
41			
42			
43			
44			
45			
46			

1			
2			
3			
4			
5	X69	Liver function	1, Less than 3 times upper limit of normal range of liver function tests (ULN of
6			LFTs); 2, 3~5 times ULN of LFTs; 3, More than 5 times ULN of LFTs
7			
8	X70	Total bilirubin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
9	X71	Lactic dehydrogenase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
10	X72	Creatine kinase	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
11	X73	White blood cell	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
12	X74	Neutrophil granulocyte	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
13	X75	Lymphocyte percentage	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
14	X76	Monocyte percentage	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
15	X77	Eosinophils	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
16	X78	Red blood cell	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
17	X79	Hemoglobin	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
18	X80	Platelet count	1, Below the normal range; 2, Within the normal range; 3, Above the normal range
19	X81	Hypersensitive C-reactive protein	0, Within the normal range; 1, Above the normal range
20	X82	Pre-treatment indicators of	0, Within the normal range; 1, Above the normal range
21		carcinoma	
22	X83	Pre-treatment serum levels	0, Within the normal range; 1, Above the normal range
23			
24			
25			
26			
27			
28			
29			
30			
31			
32			
33			
34			
35			
36			
37			
38			
39			
40			
41			
42			
43			
44			
45			
46			

**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.

**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	<b>0.802±0.157</b>	0.796-0.808	0.749±0.237	0.740-0.759	<b>0.757±0.189</b>	0.750-0.764
	RF improve filling	<b>0.887±0.094</b>	0.883-0.890	<b>0.832±0.096</b>	0.828-0.835	0.799±0.158	0.793-0.806	<b>0.751±0.240</b>	0.742-0.760	0.757±0.191	0.749-0.764
	<i>p</i> value	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>	
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	<b>0.923±0.063</b>	0.920-0.925	0.858±0.085	0.854-0.861	<b>0.849±0.079</b>	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	<b>0.859±0.085</b>	0.855-0.862	0.841±0.074	0.837-0.844	<b>0.885±0.130</b>	0.879-0.890	<b>0.859±0.093</b>	0.855-0.863
	<i>p</i> value	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>	
Variable selection											

1												
2												
3												
4												
5	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	
6												
7	Lasso selection	<b>0.889±0.089</b>	0.886-0.892	<b>0.835±0.090</b>	0.832-0.838	<b>0.801±0.165</b>	0.796-0.807	<b>0.751±0.240</b>	0.743-0.759	<b>0.758±0.196</b>	0.752-0.765	
8												
9	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	
10												
11	<i>p</i> value	<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>
12	machine											
13	learning											
14	algorithms											
15												
16												
17	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	
18												
19	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	
20												
21	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	
22												
23	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	
24												
25	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	
26												
27	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	
28												
29	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	
30												
31	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	
32												
33	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	
34												
35	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	
36												
37	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	
38												
39	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	
40												
41												
42												
43												
44												
45												
46												



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	<b>0.926±0.086</b>	0.919-0.933	<b>0.875±0.096</b>	0.867-0.883	<b>0.858±0.144</b>	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	<b>0.810±0.229</b>	0.792-0.828	<b>0.808±0.185</b>	0.793-0.822
<i>p</i> value		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>		<b><i>p</i>&lt;0.0001</b>

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

# 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 TRIPOD reporting 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
<b>Title</b>		
	<a href="#">#1</a> Identify the study as developing and / or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1

## Abstract

1		<a href="#">#2</a>	Provide a summary of objectives, study design, setting,	2
2				
3				
4			participants, sample size, predictors, outcome, statistical	
5				
6			analysis, results, and conclusions.	
7				
8				
9	<b>Introduction</b>			
10				
11				
12		<a href="#">#3a</a>	Explain the medical context (including whether diagnostic or	3
13				
14			prognostic) and rationale for developing or validating the	
15				
16			multivariable prediction model, including references to	
17				
18			existing models.	
19				
20				
21				
22		<a href="#">#3b</a>	Specify the objectives, including whether the study describes	4
23				
24			the development or validation of the model or both.	
25				
26				
27	<b>Methods</b>			
28				
29				
30	Source of data	<a href="#">#4a</a>	Describe the study design or source of data (e.g.,	5
31				
32			randomized trial, cohort, or registry data), separately for the	
33				
34			development and validation data sets, if applicable.	
35				
36				
37				
38	Source of data	<a href="#">#4b</a>	Specify the key study dates, including start of accrual; end of	5
39				
40			accrual; and, if applicable, end of follow-up.	
41				
42				
43	Participants	<a href="#">#5a</a>	Specify key elements of the study setting (e.g., primary care,	5
44				
45			secondary care, general population) including number and	
46				
47			location of centres.	
48				
49				
50				
51	Participants	<a href="#">#5b</a>	Describe eligibility criteria for participants.	5
52				
53				
54	Participants	<a href="#">#5c</a>	Give details of treatments received, if relevant	5
55				
56				
57				
58				
59				
60				

1	Outcome	<a href="#">#6a</a>	Clearly define the outcome that is predicted by the prediction	7
2			model, including how and when assessed.	
3				
4				
5				
6	Outcome	<a href="#">#6b</a>	Report any actions to blind assessment of the outcome to be	7
7			predicted.	
8				
9				
10				
11				
12	Predictors	<a href="#">#7a</a>	Clearly define all predictors used in developing or validating	6
13			the multivariable prediction model, including how and when	
14			they were measured	
15				
16				
17				
18				
19	Predictors	<a href="#">#7b</a>	Report any actions to blind assessment of predictors for the	6
20			outcome and other predictors.	
21				
22				
23				
24				
25	Sample size	<a href="#">#8</a>	Explain how the study size was arrived at.	5
26				
27				
28	Missing data	<a href="#">#9</a>	Describe how missing data were handled (e.g., complete-	6
29			case analysis, single imputation, multiple imputation) with	
30			details of any imputation method.	
31				
32				
33				
34				
35	Statistical	<a href="#">#10a</a>	If you are developing a prediction model describe how	6
36			predictors were handled in the analyses.	
37	analysis methods			
38				
39				
40				
41	Statistical	<a href="#">#10b</a>	If you are developing a prediction model, specify type of	7
42			model, all model-building procedures (including any	
43	analysis methods		predictor selection), and method for internal validation.	
44				
45				
46				
47				
48	Statistical	<a href="#">#10c</a>	If you are validating a prediction model, describe how the	7
49			predictions were calculated.	
50	analysis methods			
51				
52				
53				
54	Statistical	<a href="#">#10d</a>	Specify all measures used to assess model performance	7
55			and, if relevant, to compare multiple models.	
56	analysis methods			
57				
58				
59				
60				

1	Statistical	<a href="#">#10e</a>	If you are validating a prediction model, describe any model	7
2				
3	analysis methods		updating (e.g., recalibration) arising from the validation, if	
4			done	
5				
6				
7				
8				
9	Risk groups	<a href="#">#11</a>	Provide details on how risk groups were created, if done.	7
10				
11				
12	Development vs.	<a href="#">#12</a>	For validation, identify any differences from the development	7
13	validation		data in setting, eligibility criteria, outcome, and predictors.	
14				
15				
16				
17	<b>Results</b>			
18				
19				
20	Participants	<a href="#">#13a</a>	Describe the flow of participants through the study, including	8
21			the number of participants with and without the outcome	
22			and, if applicable, a summary of the follow-up time. A	
23			diagram may be helpful.	
24				
25				
26				
27				
28				
29				
30	Participants	<a href="#">#13b</a>	Describe the characteristics of the participants (basic	8
31			demographics, clinical features, available predictors),	
32			including the number of participants with missing data for	
33			predictors and outcome.	
34				
35				
36				
37				
38				
39				
40	Participants	<a href="#">#13c</a>	For validation, show a comparison with the development	8
41			data of the distribution of important variables (demographics,	
42			predictors and outcome).	
43				
44				
45				
46				
47				
48	Model	<a href="#">#14a</a>	If developing a model, specify the number of participants	9
49	development		and outcome events in each analysis.	
50				
51				
52				
53	Model	<a href="#">#14b</a>	If developing a model, report the unadjusted association, if	9
54	development		calculated between each candidate predictor and outcome.	
55				
56				
57				
58				
59				
60				

1	Model	<a href="#">#15a</a>	If developing a model, present the full prediction model to	9
2				
3	specification		allow predictions for individuals (i.e., all regression	
4			coefficients, and model intercept or baseline survival at a	
5			given time point).	
6				
7				
8				
9				
10				
11	Model	<a href="#">#15b</a>	If developing a prediction model, explain how to the use it.	9
12				
13	specification			
14				
15				
16	Model	<a href="#">#16</a>	Report performance measures (with CIs) for the prediction	9
17				
18	performance		model.	
19				
20				
21				
22	Model-updating	<a href="#">#17</a>	If validating a model, report the results from any model	9
23				
24			updating, if done (i.e., model specification, model	
25			performance).	
26				
27				
28				
29	<b>Discussion</b>			
30				
31				
32	Limitations	<a href="#">#18</a>	Discuss any limitations of the study (such as	17
33				
34			nonrepresentative sample, few events per predictor, missing	
35			data).	
36				
37				
38				
39				
40	Interpretation	<a href="#">#19a</a>	For validation, discuss the results with reference to	15
41				
42			performance in the development data, and any other	
43			validation data	
44				
45				
46				
47	Interpretation	<a href="#">#19b</a>	Give an overall interpretation of the results, considering	15
48				
49			objectives, limitations, results from similar studies, and other	
50			relevant evidence.	
51				
52				
53				
54				
55	Implications	<a href="#">#20</a>	Discuss the potential clinical use of the model and	16
56				
57			implications for future research	
58				
59				
60				

## Other information

- 1  
2  
3  
4 Supplementary [#21](#) Provide information about the availability of supplementary 18  
5  
6 information resources, such as study protocol, Web calculator, and data  
7  
8 sets.  
9  
10  
11  
12 Funding [#22](#) Give the source of funding and the role of the funders for the 18  
13  
14 present study.  
15  
16

17 None The TRIPOD checklist is distributed under the terms of the Creative Commons Attribution  
18 License CC-BY. This checklist can be completed online using <https://www.goodreports.org/>, a tool  
19  
20 made by the [EQUATOR Network](#) in collaboration with [Penelope.ai](#)  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60