Placenta and Reproductive Medicine

ORIGINAL ARTICLE

Effect of variable selection strategy on the predictive models for adverse pregnancy outcomes of pre-eclampsia: A retrospective study

Dongying Zheng[1], Xinyu Hao[2,3], Muhanmmad Khan[4], Fuli Kang[1], Fan Li[5], Timo Hamalainen[3,*], Lixia Wang[1,*]

[1]Department of Obstetrics and Genecology, Second Affiliated Hospital of Dalian Medical University, Dalian 116023, Liaoning Province, China
[2]School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116081, Liaoning Province, China
[3]Faculty of Information Technology, University of Jyvaskyla, Jyvaskyla 40014, Finland
[4]Institute of Zoology, University of the Punjab, Quaid-e-Azam campus, Lahore 54590, Pakistan
[5]Department of Obstetrics and Gynecology, Shengjing Hospital, China Medical University, Shenyang 110136, Liaoning Province, China

ABSTRACT

Objectives: The improvement of prediction for adverse pregnancy outcomes is quite essential to the women suffering from pre-eclampsia, while the collection of predictive indicators is the prerequisite. The traditional knowledge-based strategy for variable selection confronts challenge referring to dataset with high-dimensional or unfamiliar data. In this study, we employed five different automatic variable selection methods to screen out influential indicators, and evaluated the performance of constructed predictive models. **Methods:** Seven hundreds and thirty-three Han-Chinese women were enrolled and 56 clinical and laboratory variables were recorded. After grouping based on binary pregnancy outcomes, statistical description and analysis were performed. Then, utilizing forward stepwise logistic regression (FSLR) as the reference method, another four variable selection strategies were included for filtering contributing variables as the predictive subsets, respectively. Finally, the logistic regression prediction models were constructed by the five subsets and evaluated by the receiver operator characteristic curve. **Results:** The variables confirmed statistical significance between the adverse and satisfactory outcomes groups did not overlap with the variables selected by selection strategies. “Platelet” and “Creatinine clearance rate” were the most influential indicator to predict adverse maternal outcome, while “Birth weight of neonates” was the best indicator for predicting adverse neonatal outcome. In average, the predictive models for neonatal outcomes achieved better performance than models for maternal outcomes. “Mutual information” and “Recursive feature elimination” were the best strategy under current dataset and study design. **Conclusions:** Variable selection strategies may provide an alternative approach besides picking influential indicators by statistical significance. Future work will focus on applying different variable selection methods to the high-dimensional dataset, which includes novel or unfamiliar variables. This aims to identify the most appropriate collection of predictors that can enhance prediction ability and clinical decision-making.

Key words: pre-eclampsia, feature selection, variable selection, logistic regression, forward stepwise, mutual information, least absolute shrinkage and selection, recursive feature elimination, principal component analysis

***Corresponding Author:**
Timo Hämäläinen, Faculty of Information Technology, University of Jyvaskyla, P.O.Box 35, FI-40014, Finland.
E-mail: timo.t.hamalainen@jyu.fi. https://orcid.org/0000-0002-4168-9102
Lixia Wang, Department of Obstetrics and Genecology, Second Affiliated Hospital of Dalian Medical University, No.467 Zhongshan Road, Shahekou District, Dalian 116024, Liaoning Province, China. E-mail: wanglixia691130@163.com. https://orcid.org/0009-0006-7202-3940
Received: 16 January 2024; Revised: 20 January 2024; Accepted: 24 January 2024
https://doi.org/10.54844/prm.2024.0318

INTRODUCTION

Pre-eclampsia (PE) is a leading cause of neonatal and maternal mortality and morbidity that complicates approximately 2%-8% of all pregnancies worldwide.[1] As the precise pathogenesis of PE is not completely explored, early prediction is still challenging in clinical practice. The proposed prediction scheme for PE should be serial short-term predictive models, which can predict the potential risks before, during and after pregnancy, to facilitate individual surveillance and intervention.

Logistic regression model, as one of the most popular classification methods, the predictive performance is influenced by a set of selected influential variables.[2] Before the development of modeling, variable selection is an essential procedure in order to (1) avoid the high dimensionality which may lead to computational complexity and poor performance of the model; (2) provide a better understanding of the causal relationship between predictive outcomes and response variables;[3] suggest a cost-effective monitoring regarding these important variables. (3) Stepwise logistic regression, for instance, is an automatic variable selection method applied to logistic regression widely.

As machine learning continues to flourish and evolve in medicine rapidly to provide important contributions, feature selection, an alternative expression of variable selection, is indispensable component of the learning process, dealing with the challenge of high-dimensional data. Nowadays, many feature selection methods are applied to structured medical records as effective pre-processing procedure to eliminate redundant variables for the development of predictive models. However, there is no so-called "best strategy", only a good method for a specific problem setting.[4] Therefore, the purpose of this study is to investigate and better understand the performance of each variable selection method when it is applied to current dataset, accumulating more experience for selecting candidates.

In this study, a dataset consisting of women with a prior diagnosis of PE was employed, and clinical and laboratory variables were collected. By selecting the influential variables with different selection strategies, logistic regression models were established and the predictive accuracy for pregnancy outcomes was evaluated, aiming for providing the efficient model created by prior combination of accessible predictors. The application of variable selection methods may promise the achievement of retrieving reliable variables, especially for low resource settings.

Healthcare practitioners may not be familiar with many of the concepts introduced in this paper. With the basic level of statistical understanding, clinicians interpret the latest clinical information, it may be a meaningful attempt to access a basic understanding of machine learning which may assist for developing decision support systems for patient benefit.

METHODS

The flow chart of this research can be seen in Figure 1A.

Study population

A retrospective study was conducted consisting of women admitted to two tertiary care hospitals with delivery service in China, Shengjing Hospital of China Medical University (CMU), and Second Affiliated Hospital of Dalian Medical University (DMU), from January 1, 2007 to December 31, 2017. Medical records were reviewed and 733 Han-Chinese women diagnosed with pre-eclampsia with definite records of pregnancy outcomes were enrolled. The diagnostic criteria for "pre-eclampsia" were based on the American College of Obstetricians and Gynecologists practice bulletin.[5] This project was approved by the Ethics Committee of Shengjing Hospital of China Medical University (No.2013PS68K) and Second Affiliated Hospital of Dalian Medical University (No.2022033). This research had an exemption from informed consent.

The exclusion criteria applied were as follows: (1) uncertainty regarding the last menstrual period (LMP) and unwillingness to undergo an ultrasound scan before 14 weeks of gestation; (2) known major fetal anomaly or abnormal karyotype; (3) multiple gestation; (4) presence of cardiovascular, respiratory, hepatic, renal, immune system, hematological system diseases, acute infectious disease, major uterine anomaly, cervical cerclage, or malignant tumor; and (5) missing data rates exceeding 50% for analyzed variables.[6]

Grouping

733 women were categorized separately on the basis of maternal or neonatal adverse outcomes. (1) Grouped based on maternal outcomes: the adverse maternal outcomes group (A-M, N = 182) and the satisfactory group (S-M, N = 551); (2) Grouped based on neonatal outcomes: the adverse neonatal outcomes group (A-N, N = 423) and the satisfactory group (S-N, N = 310). The adverse maternal and neonatal outcomes regarding pre-eclampsia were identified according to the international consensus,[7] comprising 14 maternal and 8 neonatal outcomes, the detailed list of outcomes can be seen in Table 1, 2 of this reference.

Collection of variables

All maternal variables were collected within 24 h of admission, neonatal variables were recorded within 24 h

Table 1: Variables with statistical significance between the adverse maternal outcomes group and the control

Variables	A-M group (*N* = 182)	S-M group (*N* = 551)	*P* value
Gravidity	2 (1-3)	2 (1-3)	0.041
Gestational age (weeks)	33.2 (30.4-36.4)	36.9 (34.3-38.6)	<0.001
Delivery mode			<0.001
vaginal delivery	2 (1.1%)	57 (10.3%)	
forceps delivery	1 (0.5%)	2 (0.4%)	
cesarean section	146 (80.2%)	454 (82.4%)	
2nd-trimester labor induction	20 (11.0%)	30 (5.4%)	
stillbirth delivery	13 (7.1%)	8 (1.5%)	
Low birth weight	75 (41.2%)	178 (32.3%)	0.028
Birth weight of neonates (g)	1894.9 ± 975.8	2520.1 ± 993.7	<0.001
Maternal body mass index	28.8 ± 4.0	30.7 ± 4.5	<0.001
Systolic pressure (mmHg)	154.8 ± 28.4	148.4 ± 21.1	0.005
Diastolic pressure (mmHg)	100.3 ± 20.8	95.7 ± 15.0	0.006
Leukocyte ($\times 10^9$ /L)	11.56 ± 7.18	9.61 ± 3.17	<0.001
Neutrophil ($\times 10^9$/L)	27.93 (7.35-75.78)	12.76 (6.14-70.99)	0.010
Platelet ($\times 10^9$/L)	149.7 ± 69.3	194.4 ± 64.7	<0.001
APTT (s)	31.70 ± 9.16	29.89 ± 5.00	0.012
fibrinogen (g/L)	4.00 ± 1.14	4.36 ± 1.60	0.005
ALT (U/L)	22 (16.6-35.3)	17 (12-24)	<0.001
AST (U/L)	21.9 (14-34.5)	17 (12-24)	<0.001
Total protein (g/L)	53.7 ± 7.3	55.8 ± 7.0	<0.001
Albumin (g/L)	28.9 ± 4.4	30.4 ± 4.7	<0.001
Urea (mmol/L)	5.81 ± 2.60	4.43 ± 2.46	<0.001
Creatinine (μmol/L)	71.0 ± 27.1	56.9 ± 14.8	<0.001
Creatinine clearance rate	141.8 ± 53.8	176.5 ± 59.4	<0.001
Uric acid (μmol/L)	418.0 ± 104.5	374.6 ± 99.9	<0.001
Serum sodium (mmol/L)	134.8 ± 14.4	137.2 ± 2.5	0.026
Serum calcium (mmol/L)	1.97 ± 0.19	2.06 ± 0.17	<0.001
Serum phosphorus (mmol/L)	1.37 ± 0.25	1.30 ± 0.24	0.001
Urine pH	6.09 ± 0.64	6.26 ± 0.68	0.002
Spot urine protein			<0.001
negative	5 (2.7%)	106 (19.2%)	
±	10 (5.5%)	54 (9.8%)	
+	18 (9.9%)	106 (19.2%)	
++	53 (29.1%)	124 (22.5%)	
+++	72 (39.6%)	120 (21.8%)	
++++	24 (13.2%)	41 (7.4%)	
Urine glucose			0.017
negative	161 (88.5%)	519 (94.2%)	
±	16 (8.8%)	14 (2.5%)	
+	2 (1.1%)	10 (1.8%)	
++	3 (1.6%)	5 (0.9%)	
+++	0	2 (0.4%)	
++++	0	1 (0.2%)	
Urine Ketone			0.039
negative	171 (94.0%)	490 (88.9%)	
±	4 (2.2%)	20 (3.6%)	
+	2 (1.1%)	6 (1.1%)	
++	3 (1.6%)	20 (3.6%)	
+++	2 (1.1%)	9 (1.6%)	
++++	0	6 (1.1%)	
Urinary casts	1.7 (0.8-4.4)	1.3 (0.1-3.7)	0.009
24-hour urinary protein (mg)	3929.2 (1809.0-9380.0)	1860.0 (366.0-5716.3)	<0.001
Cholesterol (mmol/L)	7.08 ± 2.13	6.70 ± 2.07	0.031

APTT: activated partial thromboplastin time; ALT: alanine aminotransferase; AST: aspartate aminotransferase.

Table 2: Variables with statistical significance between the adverse neonatal outcomes group and the control

Variables	A-N group (*N* = 423)	S-N group (*N* = 310)	*P* value
Gravidity	2 (1-3)	2 (1-2)	0.001
Parity	0 (0-1)	0 (0-1)	0.006
Thyroid disease history	48 (11.3%)	15 (4.8%)	0.002
Gestational age (weeks)	32.9 ± 4.0	38.3 ± 1.9	<0.001
Delivery mode			<0.001
vaginal delivery	14 (3.3%)	45 (14.5%)	
forceps delivery	0	3 (1.0%)	
cesarean section	338 (79.9%)	262 (84.5%)	
2nd-trimester labor induction	50 (11.8%)	0	
stillbirth delivery	21 (5.0%)	0	
Birth weight of neonates (g)	1736.3 ± 777.2	3247.9 ± 627.4	<0.001
Low birth weight	219 (86.6%)	34 (13.4%)	<0.001
Maternal body mass index	29.5 ± 3.9	31.1 ± 5.1	<0.001
Systolic pressure (mmHg)	152.8 ± 25.0	146.3 ± 20.4	<0.001
Diastolic pressure (mmHg)	98.4 ± 18.3	94.8 ± 14.0	0.002
Leukocyte ($\times10^9$/L)	10.58 ± 3.76	9.48 ± 5.36	0.001
Neutrophil ($\times10^9$/L)	63.00 (8.44-74.93)	7.18 (5.52-51.60)	<0.001
Hemoglobin (g/L)	122.5 ± 22.4	118.3 ± 15.1	0.002
Hematokrit (%)	36.89 ± 6.85	35.97 ± 4.81	0.034
Platelet ($\times10^9$/L)	172.24 ± 75.01	197.88 ± 55.31	<0.001
PT (s)	10.66 ± 1.45	11.83 ± 7.92	0.003
fibrinogen (g/L)	4.11 ± 1.42	4.55 ± 1.59	<0.001
ALT (U/L)	21.0 (16.0-32.0)	14.0 (10.0-20.0)	<0.001
AST (U/L)	18.0 (12.0-30.0)	17.0 (12.0-22.1)	0.011
Total protein (g/L)	53.13 ± 7.29	58.15 ± 6.24	<0.001
Albumin (g/L)	28.62 ± 4.57	31.90 ± 4.19	<0.001
Globulin (g/L)	24.56 ± 5.16	27.81 ± 28.53	0.022
Urea (mmol/L)	5.43 ± 3.02	3.93 ± 1.38	<0.001
Creatinine (μmol/L)	64.90 ± 22.16	54.35 ± 12.29	<0.001
Creatinine clearance rate	151.76 ± 54.15	185.90 ± 64.35	<0.001
Uric acid (μmol/L)	407.77 ± 106.88	349.90 ± 89.27	<0.001
Serum sodium (mmol/L)	136.17 ± 9.67	137.29 ± 2.76	0.047
Serum calcium (mmol/L)	2.01 ± 0.20	2.07 ± 0.15	<0.001
Serum phosphorus (mmol/L)	1.36 ± 0.26	1.26 ± 0.20	<0.001
Urine leukocytes count	22.24 (2.10-65.05)	2.78 (1.00-15.00)	<0.001
Spot urine protein			<0.001
negative	25 (5.9%)	83 (26.8%)	
±	20 (4.7%)	43 (13.9%)	
+	53 (12.5%)	72 (23.2%)	
++	125 (29.6%)	57 (18.4%)	
+++	152 (35.9%)	42 (13.5%)	
++++	48 (11.3%)	13 (4.2%)	
Urine erythrocytes count	17.79 (6.67-38.92)	1.98 (0.00-11.12)	<0.001
Urine ketone			0.026
negative	389 (92.0%)	270 (87.1%)	
±	13 (3.1%)	12 (3.9%)	
+	4 (0.9%)	4 (1.3%)	
++	14 (3.3%)	12 (3.8%)	
+++	1 (0.2%)	9 (2.9%)	
++++	2 (0.5)	3 (0.9%)	
Urinary casts	1.97 (0.90-5.12)	0.88 (0.00-2.36)	<0.001
24-hour urinary protein (mg)	5520.0 (2063.8-9820.0)	675.6 (236.6-1984.5)	<0.001
Cholesterol (mmol/L)	7.12 ± 2.46	6.37 ± 1.28	<0.001
Triglyceride (mmol/L)	4.57 ± 2.94	3.89 ± 1.32	<0.001
Amniotic fluid index (cm)	5.66 ± 3.06	8.27 ± 3.93	<0.001

PT: prothrombin time; ALT: alanine aminotransferase; AST: aspartate aminotransferase

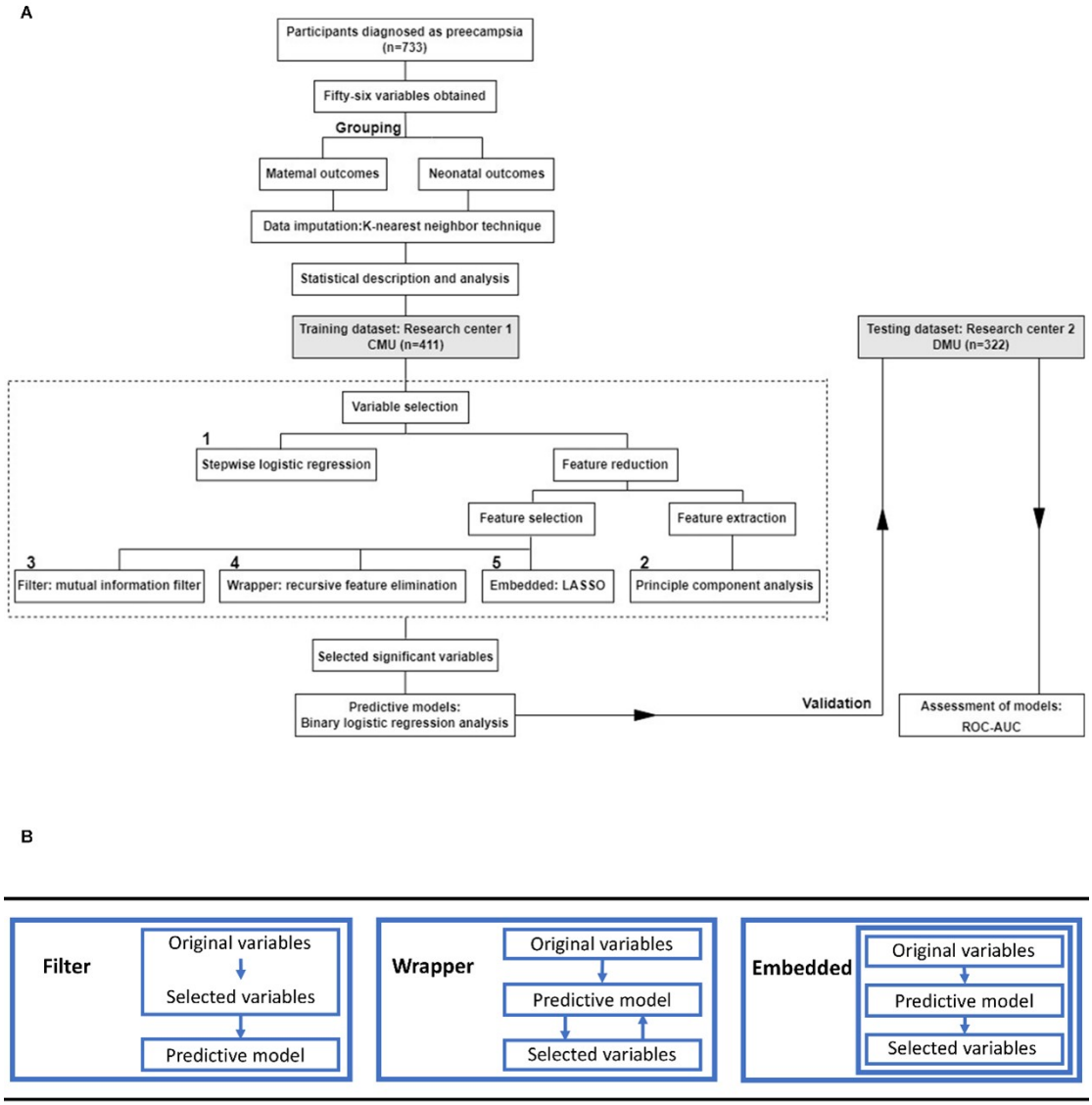

Figure 1. The flow chart of the study and the illustration of three variable selection strategies. (A) The flow chart of this study; (B) The illustrative images of three variable selection strategies: filter, wrapper and embedded.

after delivery, and biological samples were analyzed in the laboratories of two hospitals. Gestational ages were confirmed by ultrasonic examination before 14 gestational weeks.

Maternal variables included (1) Basic information (14 variables): age, gravidity, parity, weight, height, body mass index (BMI), systolic pressure, diastolic pressure, history of chronic hypertension, *in vitro* fertilization and embryo transfer (IVF-ET), advanced maternal age (> 35 years), scarred uterus, pregestational or gestational diabetes mellitus, and thyroid disease; (2) Laboratory test indexes (37 variables): 6 variables from routine blood tests, 4 variables from coagulation tests, 5 variables from liver function tests, 4 variables from kidney function tests, fasting blood glucose levels, 5 variables related to serum ion concentrations, 8 variables from routine urine dipstick testing, 24-hour urinary protein levels, 2 lipid profile-related parameters, and amniotic fluid index (detailed in Supplementary 1).

Neonatal variables included the following (5 variables): gender of neonates, low birth weight status, birth weight of neonates, gestational age at delivery, and mode of delivery.

The "creatinine clearance rates" were calculated based on the Cockcroft-Gault equation. [8]

Random missing data were inevitable in our retrospective study to threat the validation of results unnecessarily. We performed imputation method for missing value before

any subsequent analysis. K-nearest neighbor intelligent imputation was the preferential technique. This algorithm can investigate the relationships between values, and the missing values can be approximated by the "k" neighbors that are closest to them.[9] The missing data rate was calculated after data imputation.

After statistical analysis, all the variables were standardized ranging from 0 to 1. The same order of magnitude ensures the weakness or even elimination of disturbance factors, which guarantees the comparability between different variables and improves the accuracy of predictive modeling.

Selection of predictive variables

Variable selection was conducted using five distinct strategies as a pre-processing step for subsequent development of predictive models.

Stepwise logistic regression analysis

Stepwise logistic regression analysis includes three variants: stepwise forward selection, stepwise backward elimination, and stepwise combined with forward selection and backward elimination, while the last one holds the merits of forward selection and backward elimination.[10–12] The Statistical Package for the Social Sciences (SPSS) provides computing packages for either forward or backward stepwise procedure. Forward stepwise logistic regression (FSLR) was applied to the current dataset with default entry criterion.

Dimensionality reduction

High-dimensional data cause computational complexity and "curse of dimensionality". Dimensionality reduction convert the higher dimension data space into lower one, discarding redundant features. It includes two mainstream methods: feature extraction and feature selection. Feature extraction transforms the original features of data to construct lower-dimension feature spaces, while feature selection can preserve the critical information of data.[13]

Principal component analysis (PCA) is one of the most common feature extraction techniques applied. By identifying the patterns of variables to highlight their similarities and differences, this feature transformation technique converts the high-dimensional correlated variables into linear uncorrelated variables. These new independent variables derived from original variables are known as principal components (PC). The PCs holding maximum information are reserved for efficient calculation, with the properties of collecting highly correlated variables within each component and being uncorrelated with each other.[14–15]

Feature selection methods

Three major feature selection approaches can be distinguished as: filters, wrappers and embedded methods. (Figure 1B) To ensure clarity for readers with a background in medicine, the term 'variable selection' was employed instead of 'feature selection' in the following section to avoid any potential confusion.

Filters

Filters rely on the general characteristics of training data and carry out the variable selection process as a pre-processing step with independence of the modeling algorithm.[4] In this study, the proposed filter variable selection method is "mutual information (MI)".

Mutual information filter is a share the least filter. With the measurement of information amount that one random variable contains about another, the MI between two variables indicates the reduction in uncertainty of one due to the knowledge of the other.[16] By calculating the MI between candidate subsets and the outcomes of classification, this method was applied to the field of variable selection. The component variables in the selected subset share the least redundancy between each other, while demonstrate greatest correlation with the outcome.[17]

Wrappers

The wrapper methods interact with the machine learning models. All the variables are considered as a whole, each possible combination of variables is evaluated by using learning model as the performance evaluator, and critical variables are selected based on the success of the selection process. Therefore, wrappers guarantee the accurate prediction result than the filter method, but consume much more time than filters in general, consequently.[18]

Recursive feature elimination (RFE) is introduced as a wrapper. Initially, the original variables are utilized to develop the learning model (the proposed model is random forest algorithm in this study). After the accuracy scores are calculated for each variable, the variables with less scores are removed from the list and a new subset of efficient variables is constructed. The model is constructed again and the process is repeated until there is no more variable which can be excluded, or the defined number of variables is reached. Finally, the subset with advanced performance is the optimized collection of variables.[18–19]

Embedded methods

Embedded method selects variables as part of the model construction, the selection process is integrated with the modeling algorithm, indicating the variable selection is completed during the training of the model.[19]

Least absolute shrinkage and selection (LASSO) regression is an efficient embedded variable selection method. LASSO shrinks the regression coefficients of each variable by a tuning parameter, which is also known as penalty. When the parameter is sufficiently large, the absolute size of coefficients is forced to shrink. By setting as many coefficients as possible to zero, the penalized dataset is regressed with as few variables as possible. With the controlled penalizing coefficient, LASSO facilitates the desirable variables.[20–22]

As the number of variables selected by the FSLR method was determined automatically, the number of variables selected by the other four strategies was fixed based on the determination made by FSLR. This approach was adopted to facilitate a comparative analysis of the efficiency exhibited by the following models.

Predictive models

After the scheme of variable selection was employed to remove spurious variables, logistic regression analysis was performed for the development of prediction models. Binary logistic regression analysis is a non-linear regression technique that assumes the expected probability of a binary outcome. The regression coefficients quantify the contribution of variables to the probability. Comparisons were expressed as odds ratios (OR) and 95% confidence intervals (95% CI), the variables with higher odds ratios were considered to have more significant pattern changes.[23] Model calibration was assessed *via* the Hosmer-Lemeshow goodness-of-fit test.

The quality of the models was assessed by the receiver operator characteristic (ROC) curve. Over a series of thresholds, the discriminatory power of the models can be determined by the area under the curve (AUC).

The dataset from CMU (n = 411) was utilized as the "Training dataset" for model validation in our research, and procedures of "predictive variable selection" and "construction of logistic regression models" were conducted based on this dataset. Subsequently, the constructed models were applied to the dataset from DMU (n = 322), which served as the "Testing dataset". ROC curves were generated using this testing dataset, and corresponding AUC values were calculated.

For all logistic regression models, the technique we optimized to correct the bias during discriminative process was "ten-fold cross-validation". Initially, the Dataset are divided into ten equal size subsets randomly; and then, seven subsets are utilized for training and three remaining subsets are utilized for testing in each iteration; after ten iterations were performed, each subset can be used as a testing set in a rotatory manner. The final performance of models is calculated as the average of all the iterations.[24]

Statistical description and analysis

The normality of distribution was analyzed by the Shapiro-Wilk test for continuous variables. Intergroup comparisons between continuous variables with normal distributions were performed by Student's t-test and presented as mean ± standard deviation. Continuous variables with skewed distributions were compared using the Mann-Whitney U test and described as median (interquartile range). Categorical variables were analyzed by Chi-square test. Ordinal variables were compared by the Mann-Whitney U test. A probability level of P-value < 0.05 was taken as statistically significant.

All analyses were performed by: Python language version 3.6.9; SPSS version 26 (IBM Corp., Armonk, NY, USA); GraphPad Prism 6.01 (GraphPad Software, San Diego, CA, United States).

Ethical issues

Ethical approval was obtained from the Medical Ethics Committee of Second Affiliated Hospital of Dalian Medical University (2022-033) dated 28-04-2022, and the Shengjing Hospital of China Medical University (2013PS68K) dated 04-03-2013. All procedures adhered to the ethical standards with the principles of the Declaration of Helsinki. The requirement for informed consent was waived off for this retrospective and observational study. Personal information of the participants was shielded before any analysis.

RESULTS

Population characteristics

A total of 733 pregnant women with documented pregnancy outcomes were included in this study. Fifty-six clinical and laboratory variables were extracted from medical records. Missing data rate can be seen in Supplementary 1. There were 182 participants (24.8%) who met the criteria of adverse maternal outcomes, while 423 participants (57.7%) involved in adverse neonatal outcomes. Among 56 variables, thirty-one variables confirmed significantly statistical differences when compared A-M and S-M (Table 1), and 16 variables demonstrated highly statistical significance (P < 0.001). While there were 38 variables confirmed statistical differences when compared A-N and S-N (Table 2), and 26 variables demonstrated highly statistical significance.

Variables selected by different strategies

Five different variable selection strategies were applied to identify influential factors that predicted the adverse maternal or neonatal outcomes. As the number of selected variables was automatically determined by stepwise logistic regression, to access the impartial evaluation, the size of selected variable collection for each strategy was proposed to be equal to the number of variables retrieved by FSLR. Therefore, seven variables

were selected as the contributing factors to predict adverse maternal outcomes (Table 3), and eleven variables to predict adverse neonatal outcomes (Table 4). As they were displayed in two Tables, the variables selected by different strategies varied both in content and sequence.

Referring to the selected variables employed to predict adverse maternal outcomes, "Platelet" and "Creatinine clearance rate" were the two variables voted by all strategies, indicating their unique roles in the prediction of maternal outcomes. Five variables demonstrated no statistical difference between the A-M and S-M groups, when screened out by different selection strategies: "Serum chloride" and "Prothrombin time" were selected by FSLR, "Body mass index" by MI, "Urine erythrocytes count" and "Urine leukocytes count" by both MI and RFE, while variables selected by PCA and LASSO are all statistically different between the two groups.

As to the variables selected to predict neonatal outcomes, "Birth weight of neonates" was the only variable approved by all the strategies, which is easy to be understood. Meanwhile, "24-hour urinary protein", "Uric acid" and "Creatinine clearance rate" were the three variables selected by four variable selection methods out of five, all of which are renal function assessment indexes.

There were six variables screened out by different selection strategies as influential indicators demonstrating no statistical difference between the A-*N* and S-*N* groups. "Urine pH", "history of Pregestational or gestational diabetes", and "Maternal height" by FSLR; "Thrombin time" by MI; "Urine specific gravity" and "Serum potassium" by LASSO. It can be figured out that, based on our current dataset, there is no satisfactory indicator acknowledged by all the retrieving strategies to predict the neonatal outcomes in advance.

The Supplementary 2 presented the relevance of each selected variable to adverse outcomes as a score, but different variable selection algorithms resulted in varying magnitudes of scores across variables. (The variables selected by FSLR were not included as they were ranked as their sequential entry to the regression model.)

It is interesting that the PCA never singled out a variable which demonstrated no statistical significance. Then, different subsets of candidates entered into the binary logistic regression predictive models subsequently.

Evaluation of predictive models developed by different collections of variables

All variables entered into the logistic regression had a variance inflation factor (VIF) of < 5, indicating a lack of multicollinearity between them. The evaluating parameters (including AUC, Sensitivity, Specificity, Positive predictive value, and Negative predictive value) of predictive models for adverse pregnancy outcomes appear in Table 5. The overall level of AUC values of adverse neonatal outcomes predicting models is higher than that of adverse maternal outcomes (*Figure 2A-B*), which can be explained by the closer correlation between the selected variables and the neonatal outcomes.

The LR-MI predictive model exhibited the highest levels of AUC values (AUC = 0.824) for maternal adverse outcomes, while the LR-PCA model failed to achieve its optimal performance. In predicting adverse neonatal outcomes, the LR-RFE model ranked first (AUC = 0.842), and the AUC of LR-MI was 0.819, still demonstrating a strong predictive ability compared to other models. In addition, the LR-PCA model underperformed once again with the lowest AUC values.

As the variables selected by FSLR were listed according to their sequential entry to the model, in order to display the relevance of variables to adverse maternal outcomes, the OR and 95% CI were shown in *Table 6* and the corresponding forest plot was displayed in Figure 2C.

In the forest plot, the horizontal axis represents the OR and 95% CI, while the vertical dotted line positioned at a value of 1 is referred to as the "line of null effect". On the left side of this line, the protective variables such as "creatinine clearance rate", "serum chloride", and "platelet" were observed. Increasing their values effectively reduced adverse outcomes. Conversely, "leukocyte" and "creatinine" are identified as risk variables since an increase in their values elevates the probability of adverse outcomes. The OR value of the variable "Birth weight of neonates" is equal to 1, indicating no significant impact on outcomes. This can be attributed to the negligible effect observed when altering the birth weight of neonates by a mere gram.

DISCUSSION

Currently, machine learning is being applied in the field of preeclampsia across three main domains:(1) Predicting the onset of pre-eclampsia prior to any discernible symptoms;[25–27] (2) Predicting adverse neonatal and maternal outcomes associated with evident or suspected pre-eclampsia;[28–30] (3) Identifying potentially highly predictive variables through diverse datasets.[31] It is evident that numerous predictive models constructed by machine learning algorithms have been extensively reported worldwide; however, the identification of predictive variables remains inadequate.

Typically, the selection of variables for a regression predictive model is guided by hypotheses that have been formulated based on theoretical reasoning or previous

Table 3: List of variables selected by different variable selection strategies for predicting adverse maternal outcomes

Ranking	FSLR	PCA	MI	RFE	LASSO
1	Creatinine	Creatinine	Creatinine clearance rate*	24-hour urinary protein	Platelet*
2	Platelet*	Creatinine clearance rate*	Platelet*	Birth weight of neonates	Creatinine clearance rate*
3	Leukocyte	Platelet*	Body mass index	Urine erythrocytes count	Leukocyte
4	Creatinine clearance rate*	Birth weight of neonates	Urinary Casts	Platelet*	Creatinine
5	Serum chloride	Urea	24-hour urinary protein	Creatinine clearance rate*	Neutrophil
6	Prothrombin time	Gestational age	Urine erythrocytes count	Urine leukocytes count	Birth weight of neonates
7	Birth weight of neonates	Uric Acid	Urine leukocytes count	alanine aminotransferase	Diastolic pressure after admission

*represent the variable that has been voted by all strategies. The variables selected by forward stepwise logistic regression (FSLR) were ranked based on their sequential entry into the regression model, while the ranking of the variables screened by the other four strategies was determined based on their relevance to disease occurrence. PCA: principal component analysis; MI: mutual information; RFE: recursive feature elimination; LASSO: least absolute shrinkage and selection.

Table 4: List of variables selected by different variable selection strategies for predicting adverse neonatal outcomes

Ranking	FSLR	PCA	MI	RFE	LASSO
1	Birth weight of neonates*	Birth weight of neonates*	Birth weight of neonates*	24-hour urinary protein#	Birth weight of neonates*
2	Spot urine protein	Low birth weight	Gestational age	Birth weight of neonates*	Gestational age
3	Delivery mode	Gestational age	24-hour urinary protein#	Urine erythrocytes count	Amniotic fluid index
4	Amniotic fluid index	Spot Urine protein	Neutrophil	Urine leukocytes count	Neutrophil
5	Urine pH	24-hour urinary protein#	Creatinine clearance rate#	AST	Urine specific gravity
6	Low birth weight	Total protein	Hemoglobin	Uric acid#	24-hour urinary protein#
7	Fibrinogen	Albumin	Low birth weight	ALT	Prothrombin time
8	Pregestational or gestational diabetes	Creatinine clearance rate#	Uric acid#	Creatinine clearance rate#	Creatinine clearance rate#
9	Urinary Casts	Uric acid#	Platelet	Platelet	AST
10	Maternal height	Urea	Thrombin time	Creatinine	Uric acid#
11	Albumin	Creatinine	Leukocyte	Low birth weight	Serum potassium

The variables screened out by forward stepwise logistic regression (FSLR) were ranked as their sequential entry to the regression model. *indicates the variable voted by all the strategies. #were selected by the four strategies. ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; PCA: principal component analysis; MI: mutual information; RFE: recursive feature elimination; LASSO: least absolute shrinkage and selection.

Table 5: Evaluative criteria of different models for predicting adverse pregnancy outcomes

	AUC	SEN	SPE	PPV	NPV
Adverse maternal outcomes					
LR- Stepwise	0.780	0.396	0.952	0.724	0.832
LR- PCA	0.667	0.526	0.910	0.588	0.888
LR- MI	0.824	0.579	0.923	0.647	0.900
LR- RFE	0.752	0.526	0.936	0.667	0.890
LR- LASSO	0.708	0.396	0.952	0.724	0.832
Adverse neonatal outcomes					
LR- Stepwise	0.780	0.396	0.952	0.724	0.832
LR- PCA	0.767	0.679	0.913	0.760	0.875
LR- MI	0.819	0.536	0.913	0.714	0.829
LR- RFE	0.842	0.714	0.870	0.690	0.882
LR- LASSO	0.791	0.429	0.913	0.667	0.797

AUC: area under the curve; SEN: sensitivity; SPE: specificity; PPV: positive predictive value; NPV: negative predictive value; LR: logistic regression; PCA: principal component analysis; MI: mutual information; RFE: recursive feature elimination; LASSO: least absolute shrinkage and selection.

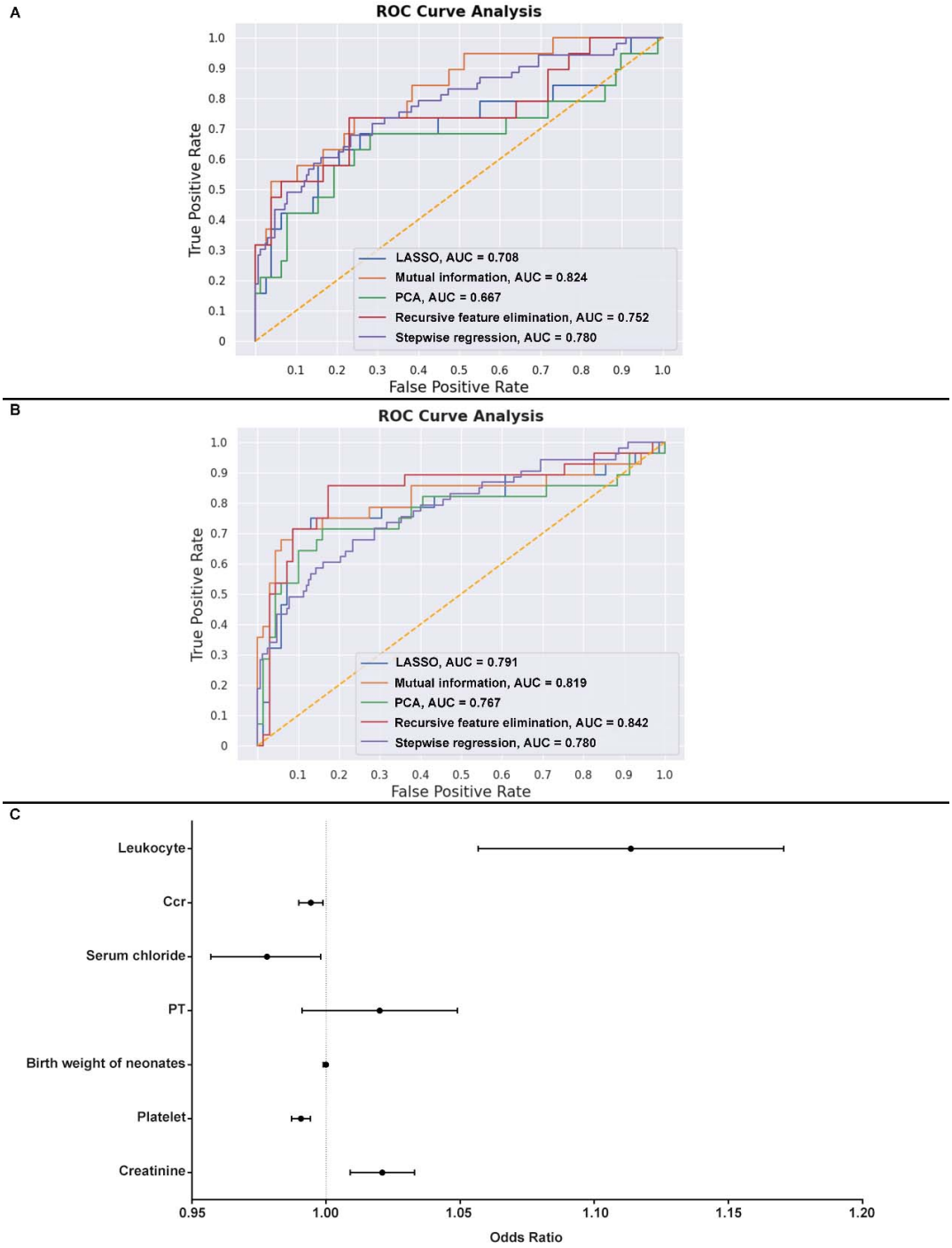

Figure 2. The receiver operator characteristic (ROC) curves of the predictive models for adverse maternal and neonatal outcomes and the forest plot of the variables selected by forward stepwise logistic regression (FSLR). (A) The AUC values of the predictive models for adverse maternal outcomes constructed by the variables selected by five different strategies; (B) The AUC values of the models for adverse neonatal outcomes; (C) The forest plot of the variables selected by FSLR, showing Odds ratio and 95% confidence intervals of the variables. Ccr: creatinine clearance rate.

empirical evidence.[32] However, this traditional opinion confronts challenge, for which high-dimensional dataset with redundant or unfamiliar variables becomes a critical issue and requires new effective, or even automatic variable selection methods. In consequence, our study was conducted to apply five different variable selection methods to screen out influential variables for predicting adverse outcomes involved pre-eclampsia, that the variables were collected at the point of adverse events, then employ FSLR as the reference model, and evaluate the performance of five methods by AUC values when they developed predictive models.

The first main finding of this study is that the variables

demonstrated statistical significance may not overlap with the variables filtered by automatic variable selection strategies. There were 31 variables confirmed statistical significance between the adverse and satisfactory maternal outcomes groups, and 38 variables between the two neonatal outcomes groups, the choice of variables which would be the true predictors probably becomes a time and energy consuming issue. As to the variables screened out by FSLR, there were only seven variables contributed to the maternal outcomes and eleven variables to the neonatal outcomes, which made the construction of predictive models more precise and explanatory. Referring to the variables filtered by selection methods, they were all reasonable to predict the adverse pregnancy outcomes according to the previous evidence, and there were also several variables confirmed no statistical significance. Meanwhile, despite FSLR, other selection strategies could calculate the predictive scores of filtered variables and rank the variables by the influential importance, while the prediction ability of variables enrolled based on statistical significance could only be valued by the odds ratio and its 95% confidence interval. Though it still needs further investigation to compare the efficiency of both statistical significancy based strategy and data-driven variable selection strategies, these automatic selection methods may provide an alternative choice, except for selecting variables only by statistical significance.

Regarding the variables selected by different variable selection strategies, both "Platelet" and "Creatinine clearance rate" were unanimously identified as predictors of adverse maternal outcomes, while four out of five strategies concurred on selecting "24-hour urinary protein," "Uric acid," and "Creatinine clearance rate" as predictors of adverse neonatal outcomes. Our findings provide evidence that targeted determinants for adverse maternal and neonatal outcomes may be associated with distinct mechanisms; for instance, severe maternal complications could involve coagulation dysfunction and renal injury,[33–34] whereas neonatal adverse outcomes might be attributed to hypoproteinemia and metabolic disorder.[34–37] Consequently, prevention and treatment measures should be tailored according to the specific situation.

The second main finding is that MI and RFE performed best under our current dataset, when predicting maternal and neonatal outcomes. Filter is an established variable selection strategy that does not rely on subsequent classification techniques and is computationally efficient. Conversely, the wrapper selects candidate variables based on the performance of classification techniques and effectively solves optimization problems, but incurs high computational costs due to evaluating and selecting all combinations of variables. Additionally, the embedded method integrates as a component of the classification algorithm, combining the strengths of both wrapper and filter.[38] The LASSO, however, did not yield models with high predictive accuracy in this study. This outcome can be attributed to the fact that the LASSO performs optimally when dealing with datasets characterized by a small sample size and a large number of variables. Under our current structural dataset which is a relatively low-dimensional issue with less irrelevant and redundant variables, the embedded approach could not sufficiently advanced for constructing satisfactory predictive model, as well as FSLR and PCA.

Concerning to the higher AUC values of neonatal outcome prediction, the variables we collected possess too direct relationship to the outcomes, like birth weight of neonates, which made the AUC values of the predictive models relatively high. While on the other hand, it indicates that there remains a lack of satisfactory variables for predicting fetal or neonatal outcomes.

Regarding the variables included by FSLR in predicting adverse maternal outcomes, we ranked these variables based on their sequential entry into the regression models. However, due to a lack of information regarding their influential weights on the outcomes, we presented the relevant OR, 95% CI, and statistical significance using an intuitive forest plot. The clinical significance of these selected variables aligns with our current knowledge as illustrated in Result 3. Furthermore, to concisely present our research findings, corresponding OR values and 95% CI for the variables selected to predict neonatal outcomes were not provided; however, interested readers can easily derive these results from our dataset when communicating with the corresponding authors.

Missing data is a commonly encountered problem in clinical research. In this study, we set the enrolled criteria of variables with missing data rate less than 50%,[39] which was still too high and may be considered as a limitation of our study. However, the missing problem may be induced at random, while it is more often that the problem is induced not at random, directly or indirectly. The lack of international or regional consensus which control the unified scheme of clinical variables obtained, as well as the emergency situation encountered more often referring to pre-eclampsia, all lead to the variables associated to rescue or surgery demonstrated lower rate of missing values, for instance of routine complete blood count or blood coagulation test. In consequence, simply excluding the variables with relatively high missing data rate may cause unforeseen biases which impact on the accuracy or validation of modeling. Meanwhile, if the missing data rate of particular variable is too high, it may indicate the difficulty of access in study population for which may

not be an ideal predictor to be included. (2) Therefore, the met criteria were set as 50%, and imputation technique was applied to avoid removing variables with relatively higher missing data rate. According to our previous study,[40] k-nearest neighbor intelligent imputation was the most effective method, and was employed in this study. After all, imputation is not the final solution, but the promotion of powerful predictive model to guarantee the establishment of international consensus and the avoidance of emergency.

Despite missing values, another limitation of our study is the tuned number of variables selected by different selection methods. As the FSLR served as the reference selection strategy, the number of filtered variables was automatically determined, while for other strategies, the parameterization of variable count was adjusted to match that synthesized by FSLR, aiming to facilitate an impartial comparison among different strategies when employing selected variables subsequently in constructing predictive models. However, it is possible to select the number of variables automatically, which can be applicable by performing cross-validation to evaluate the subsets containing different numbers of variables and to automatically select the number of variables that achieved the best mean score. (4) Therefore, the expected number of variables to predict pregnancy outcomes may prevent the other four strategies from achieving their best performance.

Future work will concentrate on the accumulation of more evidence that (1) the automatic variable selection strategies give full play to the role of screening out influential predictors to predict the outcomes of pre-eclampsia, especially in high-dimensional data, like involving variables collected before, during and after pregnancies, as well as the novel biomarkers; (2) concentrate on the investigation of matching degree between certain selection strategy and certain dataset, aiming for exploring the optimized selection method to restrict the number of predictive variables and validate their significance; (3) focus on the construction of serial short-term predictive models to predict risks in different timespan of pregnancy timely and efficiently. If so, the contributing predictive models may lead to the establishment of standard clinical protocols and assist the process of decision-making by clinical practitioners, and finally benefits to this "mother and offspring" conflict.

CONCLUSIONS

Our study showed that under the current dataset, the variables demonstrated statistical significance may not completely overlap with the variables filtered by automatic variable selection strategies, which provide an alternative choice. "Platelet count" and "Creatinine clearance rate" were unanimously identified as predictors of adverse maternal outcomes, while four out of five strategies concurred on selecting "24-hour urinary protein," "Uric acid," and "Creatinine clearance rate" as predictors of adverse neonatal outcomes. These findings suggest that distinct mechanisms may be associated with targeted determinants for adverse maternal and neonatal outcomes, highlighting the need for tailored prevention and treatment measures according to specific situation. The MI-LR and RFE-LR exhibited a higher likelihood of selecting influential variables compared to other strategies in the current dataset. However, the LASSO-LR method did not demonstrate sufficient advancement in constructing models for this relatively low-dimensional dataset with fewer irrelevant and redundant variables. Medical researchers who develop regression models for clinical prediction with high-dimensional data or limited knowledge about novel indicators could thus benefit from using automatic strategies. Future work will concentrate on applying different selection methods to high-dimensional data and constructing serial short-term predictive models to further explore the characters of these selection strategies.

DECLARATION

Supplementary materials

Supplementary materials mentioned in this article are online available at the journal's official site only.

Acknowledgement

We acknowledge the study participants.

Author contributions

Zheng DY and Hao XY devised the study plan; Li F and Kang FL helped with data acquisition; Zheng DY built the dataset, analyzed the data and drafted the article; Hao XY contributed to interpretation of data; Khan M completed language editing and revised the draft; Hamalainen T and Wang LX supervised the research. All authors read the draft manuscript and made important intellectual contributions to the final version.

Ethics approval

This project was approved by the Ethics Committee of Shengjing Hospital of China Medical University (No.2013PS68K) and Second Affiliated Hospital of Dalian Medical University (No.2022033). This research had an exemption from informed consent.

Source of funding

Not applicable.

Conflict of interest

The authors declare no competing interest.

Data availability statement

Not applicable.

REFERENCES

1. Giannakou K. Prediction of pre-eclampsia. *Obstet Med* 2021;14:220-224.
2. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. *J Thorac Dis* 2019;11:S574-S584.
3. Tsai TL, Huang MH, Lee CY, Lai WW. Data Science for Extubation Prediction and Value of Information in Surgical Intensive Care Unit. *J Clin Med* 2019;8(10):1709.
4. Bolon-Canedo V, Sanchez-Marono N, Alonso-Betanzos A. A review of feature selection methods on synthetic data. *Knowl Inf Syst* 2013;34:483-519.
5. ACOG Practice Bulletin No. 202: Gestational Hypertension and Preeclampsia. *Obstet Gynecol* 2019;133(1):1.
6. Dong W, Fong DYT, Yoon JS, *et al.* Generative adversarial networks for imputing missing data for big data clinical research. *BMC Med Res Methodol* 2021;21(1):78.
7. Duffy J, Cairns AE, Richards-Doran D, *et al.* A core outcome set for pre-eclampsia research: an international consensus development study. *BJOG* 2020;127(12):1516-1526.
8. Fernandez-Prado R, Castillo-Rodriguez E, Velez-Arribas FJ, Gracia-Iguacel C, Ortiz A. Creatinine Clearance Is Not Equal to Glomerular Filtration Rate and Cockcroft-Gault Equation Is Not Equal to CKD-EPI Collaboration Equation. *Am J Med* 2016;129(12):1259-1263.
9. Triguero I, Garcia-Gil D, Maillo J, Luengo J, Garcia S, Herrera F. Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. *Wires Data Min Knowl* 2019;9:e1289.
10. Arunajadai SG. Stepwise logistic regression. *Anesth Analg* 2009;109:285, 285-286.
11. Lee CY, Chen BS. Mutually-exclusive-and-collectively-exhaustive feature selection scheme. *Appl Soft Comput* 2018;68:961-971.
12. Pace NL. Independent predictors from stepwise logistic regression may be nothing more than publishable P values. *Anesth Analg* 2008;107(6):1775-1778.
13. Wang ZH, Liang SL, Xu LZ, Song W, Wang DX, Huang DM. Dimensionality reduction method for hyperspectral image analysis based on rough set theory. *Eur J Remote Sens* 2020;53:192-200.
14. Ray P, Reddy SS, Banerjee T. Various dimension reduction techniques for high dimensional data analysis: a review. *Artif Intell Rev* 2021;54:3473-3515.
15. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci* 2016;374(2065):20150202.
16. Battiti R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans Neural Netw* 1994;5(4):537-550.
17. Cheng J, Sun J, Yao K, Xu M, Cao Y. A variable selection method based on mutual information and variance inflation factor. *Spectrochim Acta A Mol Biomol Spectrosc* 2022;268:120652.
18. Subbiah SS, Chinnappan J. Deep learning based short term load forecasting with hybrid feature selection. *Electr Pow Syst Res* 2022;210:108065.
19. Chen Q, Meng Z, Liu X, Jin Q, Su R. Decision Variants for the Automatic Determination of Optimal Feature Subset in RF-RFE. *Genes (Basel)*. 2018;9(6):301.
20. Ueno D, Kawabe H, Yamasaki S, Demura T, Kato K. Feature selection for RNA cleavage efficiency at specific sites using the LASSO regression model in Arabidopsis thaliana. *BMC Bioinformatics*. 2021;22(1):380.
21. Zhou Y, Uddin MS, Habib T, Chi GT, Yuan KP. Feature selection in credit risk modeling: an international evidence. *Econ Res-Ekon Istraz* 2021;34:3064-3091.
22. Tibshirani R. Regression shrinkage and selection *via* the Lasso. *J R Stat Soc B.* 1996;58:267-288.
23. Tirzïte M, Bukovskis M, Strazda G, Jurka N, Taivans I. Detection of lung cancer with electronic nose and logistic regression analysis. *J Breath Res.* 2018;13(1):016006.
24. Wong TT, Yang NY. Dependency Analysis of Accuracy Estimates in k-Fold Cross Validation. *IEEE Trans Knowl Data Eng* 2017;29:2417-2427.
25. Liu M, Yang X, Chen G, Ding Y, Shi M, Sun L, *et al.* Development of a prediction model on preeclampsia using machine learning-based method: a retrospective cohort study in China. *Front Physiol* 2022;13:896969.
26. Melinte-Popescu AS, Vasilache IA, Socolov D, Melinte-Popescu M. Predictive Performance of Machine Learning-Based Methods for the Prediction of Preeclampsia-A Prospective Study. *J Clin Med* 2023;12(2):418.
27. I Maric, A Tsur, N Aghaeepour, A Montanari, DK Stevenson, GM Shaw, *et al.* Early prediction of preeclampsia *via* machine learning. *Am J Obstet Gynecol MFM* 2020;2(2):100100.
28. Espinola-Sánchez M, Sanca-Valeriano S, Campaña-Acuña A, Caballero-Alvarado J. Prediction of neonatal death in pregnant women in an intensive care unit: Application of machine learning models. *Heliyon* 2023;9(10):e20693.
29. Wang G, Zhang Y, Li S, Zhang J, Jiang D, Li X, *et al.* A Machine Learning-Based Prediction Model for Cardiovascular Risk in Women With Preeclampsia. *Front Cardiovasc Med* 2021;8:736491.
30. Villalaín C, Herraiz I, Domínguez-Del Olmo P, Angulo P, Ayala JL, Galindo A. Prediction of Delivery Within 7 Days After Diagnosis of Early Onset Preeclampsia Using Machine-Learning Models. *Front Cardiovasc Med* 2022;9:910701.
31. Hackelöer M, Schmidt L, Verlohren S. New advances in prediction and surveillance of preeclampsia: role of machine learning approaches and remote monitoring. *Arch Gynecol Obstet* 2023;308(6):1663-1677.
32. Scherr S, Zhou J. Automatically Identifying Relevant Variables for Linear Regression with the Lasso Method: A Methodological Primer for its Application with R and a Performance Contrast Simulation with Alternative Selection Strategies. *Commun Methods Meas* 2020;14:204-211.
33. Gibbins JM. Adding fuel to the flames in preeclampsia: the platelet connection. *J Thromb Haemost* 2023;21:1750-1752.
34. Piani F, Agnoletti D, Baracchi A, Scarduelli S, Verde C, Tossetta G, *et al.* Serum uric acid to creatinine ratio and risk of preeclampsia and adverse pregnancy outcomes. *J Hypertens* 2023;41(8):1333-1338.
35. Pecoraro V, Trenti T. Predictive value of serum uric acid levels for adverse maternal and perinatal outcomes in pregnant women with high blood pressure. A systematic review and meta-analysis. *Eur J Obstet Gynecol Reprod Biol* 2020;252:447-454.
36. Lei T, Qiu T, Liao W, Li K, Lai X, Huang H, *et al.* Proteinuria may be an indicator of adverse pregnancy outcomes in patients with preeclampsia: a retrospective study. *Reprod Biol Endocrinol* 2021;19(1):71.
37. Morikawa M, Mayama M, Saito Y, Nakagawa-Akabane K, Umazume T, *et al.* Severe proteinuria as a parameter of worse perinatal/neonatal outcomes in women with preeclampsia. *Pregnancy Hypertens* 2020;19:119-126.
38. Alhassan AM, Zainon WMNW. Review of Feature Selection, Dimensionality Reduction and Classification for Chronic Disease Diagnosis. *IEEE ACCESS* 2021;9:87310-87317.
39. Vanhatalo J, Li Z, Sillanpää MJ. A Gaussian process model and Bayesian variable selection for mapping function-valued quantitative traits with incomplete phenotypic data. *Bioinformatics* 2019;35(19):3684-3692.
40. Zheng D, Hao X, Khan M, Wang L, Li F, Xiang N, *et al.* Comparison of machine learning and logistic regression as predictive models for adverse maternal and neonatal outcomes of preeclampsia: A retrospective study. *Front Cardiovasc Med* 2022;9:959649.