RISK EARLY WARNING MODEL FOR FEMALE PELVIC FLOOR DYSFUNCTION, METHOD OF CONSTRUCTING IT AND SYSTEM THEREOF

Technical field

The present disclosure relates to the field of medical early warning system, and in particular, to a risk early warning model for female pelvic floor dysfunction, a method of constructing the risk early warning model, and a system for the risk early warning model.

Background

The pelvic floor structurally and functionally consists of the pelvis, muscles of the pelvic floor, nerves, ligaments and blood vessels. It cooperates with the urinary, reproductive and anal systems to mechanically support the whole pelvic organs under control of endocrine, neural and nutritional metabolism systems. Pelvic floor dysfunction (PFD) refers to abnormalities in the position and function of pelvic organs due to weakened pelvic floor support structures that result from various factors, and PFD includes pelvic organ prolapse (POP), stress urinary incontinence (SUI), sexual dysfunction (SD), and chronic pelvic pain (CPP) and the like. Globally, pelvic organ prolapse reaches an incidence of 20%, with an increasing trend year by year. A survey of the incidence of PFD among women conducted in developing countries shows that pelvic organ prolapse (POP) affects 48.3% of women. Although POP does not directly threaten life, it seriously affects patients' quality of life and physical and mental health, posing substantial burdens on individuals, families and society. Predictions indicate that by 2050, China will have 483 million elderly people aged over 60 and 108 million elderly people aged over 80. With the introduction of the three-child policy, PFD, a disease developed and exacerbated by aging, is expected to intensify the conflict between people's aspirations for quality of life and medical services.

The pelvic floor is a complex structure involving multiple tissues such as muscles, fascia, ligament, blood vessel, nerves, integrating different organs such as urinary, reproductive and anorectal organs. When human body is upright, the pelvic floor bears the pressure load of the visceral organs throughout the body and is affected by systemic circulation, respiration, nutrition and metabolic changes of the entire body. Therefore, the structural and functional status of the pelvic floor reflects the locally and dynamically conditions of entire body. Factors like pregnancy, childbirth, obesity, medications, estrogen deficiency, pelvic floor surgeries, inflammation, trauma, pelvic pressure and circulation interact intricately to cause PFD, resulting in diverse structural and functional status among various patients. PFD is a chronic and cumulative process, with the aggravated physiological deterioration of myofascial structures and the related pathological injuries over time, and is influenced by various internal and external pelvic factors, which is easily overlooked at an early stage, especially the external pelvic factors. All these circumstances hinder the effective long-term management of PFD.

Currently, the diagnosis of PFD primarily relies on clinical phenotype, medical history and relevant functional examinations, lacking functional molecular markers that can reflect the long-term pathological process for early identification, intervention and evaluation. So, it depends heavily on the experience of the clinician to finish basic pelvic floor examination, as well as choices and procedures related to surgeries. Moreover, PFD patients have the characteristics of mobility between different healthcare institutions, living at home, and dispersal, leading to discontinuity poor homogeneity, and ineffective quality control and management across institutions, information, and processing in the diagnosis and treatment. Also, the diagnosis of pelvic floor diseases is highly specialized, compounded by a shortage of specialists and the lack of simple, precise, uniform and multidimensional testing and monitoring ways in clinical practice. All these factors lead to many problems such as the absence of early warning methods, insufficient diagnostic accuracy, and uncertain efficacy, so that there is no established model for early

diagnosis and treatment of PFD.

In modern medical research, the risk early warning models for the diseases are often used to predict the future incidence of a certain disease. Specifically, statistical models are established based on the multifactorial causes of a disease, to predict the probability of future occurrence of specific events among population with certain features. In recent years, early warning models have received increasing attention and have been applied in many fields of disease research. For example, Publication No. CN111640509A discloses a method and system for predicting the risk of recurrence of cervical cancer after surgery, and Publication No. CN110200619A discloses a method and device for establishing risk early warning model for major cardiovascular diseases. At present, there are no reports on the study and use of early warning models for PFD. Therefore, exploration and establishment of risk early warning models for PFD are important measures to promote the prevention and treatment of female pelvic floor dysfunction.

SUMMARY

Disclosed herein is a risk early warning model for female pelvic floor dysfunction and a method and a system of constructing the early warning model. They can achieve screening and early warning analysis of anterior vaginal wall prolapse, uterine/fornix prolapse, posterior vaginal wall prolapse, stress urinary incontinence, vaginal relaxation degree, pelvic floor muscle strength, etc., in the female pelvic floor dysfunction (PFD), by screening the variables related to risk factors (i.e., high-risk factors) and constructing a risk early warning model for the female pelvic floor dysfunction (PFD). Additionally, disclosed herein is a computer system for applying the risk early warning model for PFD.

The present disclosure is achieved by a method of constructing a risk early warning model for female pelvic floor dysfunction, comprising the steps of:

S1: obtaining risk factors for a risk population in a region;

S2: screening correlated variables from the risk factors and constructing a risk

early warning model; and

S3: training the risk early warning model to obtain an optimized model.

In S1, the risk population in a region refers to a susceptible population in the selected area.

In S1, the risk factors comprise external-risk and internal-risk factors of pelvic floor. The external-risk factors include age, BMI, diabetes, physical labor, chronic cough, constipation, and smoking. The internal-risk factors include gravidity or parity, delivery mode, pelvic floor muscle strength, and levator ani muscle injury.

In S2, the step of screening the correlated variables from the risk factors and constructing the risk early warning model comprises the steps of:

a. screening pre-selected variables of the model by using Logistic regression method (LR) according to the risk factors;

b. screening correlated variables of the model by a support vector machine-recursive feature elimination method according to the pre-selected variables of the model; and

c. constructing a risk early warning model by Machine learning method according to the correlated variables screened.

In step a, in the Logistic regression method, the formula is:

$$\mathrm{logit(P)} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_\mathrm{p} \mathrm{X}_\mathrm{p}$$

wherein, P denotes to a conditional probability of a positive result, β denotes to a partial regression coefficient, X denotes to an independent variable (exposure factor), β_0 denotes to a constant term, "1, 2…p" denotes to the number of the partial regression coefficient and independent variable (exposure factor).

In step b, in the support vector machine-recursive feature elimination method, the formula is:

$$w = \sum_k a_k y_k x_k$$

wherein, w denotes to a feature weight, a denotes to a feature parameter, y denotes to a category label, x denotes to a training sample, k denotes to a number.

In step c, the Machine learning method comprises Logistic regression method, support vector machine method, Gaussian Naive Bayes (GNB), Random forest algorithm (RF), Gradient boosting machine (GBM), AdaBoost ensemble learning model (ADA) and multilayer perception (MLP).

In step c, outputs of the risk early warning model comprise prediction accuracy (AUC), recall rate and f1-score.

Further, in S3, a population with different levels of risk is obtained through a cohort study, and an artificial intelligence machine learning algorithm is used to train the risk early warning model to obtain an optimized model.

Disclosed herein is a risk early warning model for female pelvic floor dysfunction by using the above-mentioned method.

Disclosed herein is a computer system, comprising a memory, a processor, and a computer program storing and running on the processor, wherein the method of constructing the model is performed when the processor executes the computer program.

Terms definition

AUC stands for area under the receiver operating characteristic curve, for evaluating the prediction accuracy of classifier (prediction model).

Recall rate, also known as a recall ratio, refers to a ratio of correctly predicted positive samples out of all actual positive samples.

F1-score is calculated by arithmetic mean divided by geometric mean.

Compared with the prior art, the present disclosure has the following advantages:

The present disclosure constructs an early warning system for PFD, conducting preliminary screening for the risk of pelvic floor dysfunction among healthy people (non-pelvic floor dysfunction), at-risk population, and patient portraits, etc., and it achieves risk stratification of the population to categorize the population into high-risk and low-risk populations, facilitating subsequent auxiliary diagnosis,

comprehensive management and hierarchical diagnosis and treatment for target populations by means of various follow-up strategies.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 depicts the technical process of Embodiment 2.

Figure 2 depicts the modeling results of preliminary screening data in Embodiment 2.

Figure 3 depicts a risk early warning model of the Embodiment 2.

Figure 4 depicts an analysis flow chart of Embodiment 3.

Figure 5 depicts the corresponding curve of dataset 1- Recursive feature elimination (RFE) variable data vs AUC values in Embodiment 3.

Figure 6 depicts the corresponding curve of dataset 2-RFE variable data vs AUC values in Embodiment 3.

Figure 7 depicts the curve of dataset 1-ROC in Embodiment 3.

Figure 8 depicts SHAP plot associated to variable of dataset 1 using RF model in Embodiment 3.

Figure 9 depicts the curve of dataset 2-ROC in Embodiment 3.

Figure 10 depicts SHAP plot associated to variable of dataset 2 using RF model in Embodiment 3.

DETAILED DESCRIPTION

The purpose, the technical solution and beneficial effects of the present disclosure are described in further detail below.

It should be noted that the following detailed descriptions are exemplary and are intended to provide further illustration of the claimed invention. Unless otherwise indicated, all technical and scientific terms used herein have the same meanings as those that are commonly understood by one skilled in the art to which the present disclosure belongs.

Embodiment 1: construction of risk early warning model

In this embodiment, the risk early warning model described is mainly used for screening and early warning analyses of female pelvic floor dysfunction, which mainly include: anterior vaginal wall prolapse, uterine/fornix prolapse, posterior vaginal wall prolapse, stress urinary incontinence, vaginal relaxation degree, pelvic floor muscle strength, etc.

Firstly, risk factors such as external-risk and internal-risk factors of pelvic floor are obtained for the risk populations in a region (i.e. populations are susceptible to PFD in the selected region). The external-risk factors of pelvic floor include age, BMI, diabetes, physical labor, chronic cough, constipation, and smoking, etc. The internal-risk factors of pelvic floor include gravidity or parity, delivery mode, pelvic floor muscle strength, and levator ani muscle injury, etc.

According to the risk factors, the pre-selected variables of the model are screened by Logistic regression method and the formula is:

$$\mathrm{logit(P)} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_\mathrm{p} \mathrm{X}_\mathrm{p}$$

wherein, P denotes to a conditional probability of a positive result, β denotes to a partial regression coefficient, X denotes to an independent variable (exposure factor), β_0 denotes to a constant term, "1, 2,….p" denotes to the number of the partial regression coefficient and independent variable (exposure factor).

Further, according to the screened pre-selected variables of the model, correlated variables are screened by the support vector machine-recursive feature elimination method and the formula is:

$$w = \sum_k a_k y_k x_k$$

wherein, w denotes to the feature weight, a denotes to the feature parameter, y denotes to the category label, x denotes to the training sample, k denotes to the number.

Further, according to the screened correlated variables, the risk early warning

model is constructed by the Machine learning method, such as Logistic regression method, a support vector machine method, Gaussian Naive Bayes, Random forest algorithm, Gradient boosting machine, AdaBoost ensemble learning model and multilayer perception, etc., to obtain a prediction accuracy (AUC), a recall rate and a f1-score.

In one specific embodiment, according to the screened correlated variables, a high-risk early warning model is obtained by Random forest algorithm and the formula is:

$$VIMj = \frac{VIMj}{\sum_{i=1}^{c} VIMi}$$

Wherein, VIM denotes to a variable importance measure, j denotes to a number of the feature, i denotes to a number of Random forest tree, c denotes to an amount of Random forest tree.

Finally, the population with different levels of risk is obtained through a cohort study, and an artificial intelligence machine learning algorithm is used to train the risk early warning model to obtain an optimized model.

In one specific embodiment, a computer system comprises the above-mentioned method of construing the model, and it specifically comprises a memory, a processor, and a computer program storing and running on the processor. The method of constructing the model is performed when the processor executes the computer program.

The following examples are two typical embodiments for further illustrating the specific embodiments of the present disclosure. The protection scope of the present disclosure is not limited to the specific embodiments below.

Embodiment 2: construction of risk early warning model for analyzing vaginal relaxation

In this embodiment, for risk populations in a region, external-risk factors

(including age, BMI, diabetes, physical labor, chronic cough, constipation, and smoking, etc.) and internal-risk factors (including gravidity or parity, delivery mode, pelvic floor muscle strength /myoelectricity, and levator ani muscle injury, etc.) of the pelvic floor are collected, and the ranking of risk factors are obtained by Logistic regression method and Random forest algorithm. The risk early warning model is constructed, and then the model is optimized through an artificial intelligence machine learning and sequence verification. The specific technical process is shown in Figure 1.

An epidemiological survey for pelvic organ prolapses among female patients in medical clinics in a region shows that POP is the most common PFD and female aged 55 or older and a history of vaginal delivery are two independent risk factors for increasing the pelvic organ prolapse, as shown in Table 1 below.

Table 1 Analysis of independent risk factors for pelvic organ prolapse

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimates	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-8.0398	197. 8	0.0017	0. 9676
Age	≥55	1	0.8969	0. 2111	18. 0453	< .0001
BMI	≥25	1	0. 1603	0. 1874	0. 7312	0. 3925
Occupation	Mental worker	1	-0. 1622	0.0916	3. 1381	0.0765
Gravidity	2	1	0. 0356	0.0683	0. 2721	0. 602
Delivery mode	Vaginal delivery	1	0.4318	0. 0671	41. 457	< .0001
Diabetes	Yes	1	-0. 0357	0. 1369	0. 0679	0. 7945
Hypertension	Yes	1	0.1988	0. 226	0. 7738	0. 379
Heredity	Yes	1	-0. 1648	0. 1455	1. 2828	0. 2574
Smoking	Yes	1	-6. 539	197.8	0.0011	0. 9736
Chronic cough	Yes	1	0.271	0. 2212	1. 5008	0. 2206

Constipation	Yes	1	-0. 0673	0. 0761	0. 7818	0. 3766

Note: Estimate is Maximum Likelihood Estimates, which is derived from the likelihood function and serves as a composite indicator of sensitivity and specificity; Standard Error is defined as the standard deviation divided by the square root of the sample size (N), indicating a sampling error; Wald Chi-Square is the chi-square statistic, which is the ratio of the observed frequency to the expected frequency, indicating the correlation of the test data; “Pr > ChiSq” is the statistical significance, which is determined by the chi-square statistic, indicating whether there is a correlation between two variables.

In this embodiment, cases of the clinical service are collected from January 2019 to August 2020, with basic information such as height, weight, occupation types, profession, education, gravidity, parity, fetus number, disease history, family history of disease, etc., and test information comprising the multiple test results obtained from the tests such as Point Aa located in the midline of the anterior vaginal wall 3cm , Point Ba that represents the most distal position of the anterior vaginal wall, Point C on either the lowest edge of the cervix or the vaginal cuff, genital hiatus (gh), and pre-resting phase-test value, Type I muscle digital examination, A3 feedback, urine volume, maximum urine flow, pelvic floor muscle tone in static conditions, pelvic floor muscle tone in dynamic conditions, anterior vaginal wall prolapse, uterine/fornix prolapse, posterior vaginal wall prolapse, and the like. The data are pre-processed using data mining techniques, followed by descriptive statistics of variables, variable importance ranking, and contribution of individualized variable. When sufficient parameter adjustments are performed by training a large number of patient samples, the model is adequately parameterized to establish an artificial intelligence early warning model for severe pelvic floor dysfunction.

For example, when the data are pre-processed using data mining techniques, data are de-duplicated and cleaned to delete invalid variables and time variables, and target variables can be constructed, sorting out independent variables and target

variables for the subsequent analysis, so that one target variable correspond to multiple independent variables. When carrying out descriptive statistics of variables, the statistical test method can be used to analyze whether there is a significant correlation between the independent variables and the target variables, so as to achieve the correlation analysis of the independent variables and the target variables to obtain test statistics and statistical significance. When ranking the variable importance, the random forest method can be used to obtain the order of variable importance, so as to clarify the magnitude of correlation between independent variables and target variables to obtain the importance of the independent variables. When evaluating the contribution of individualized variable, the contribution of variables can be obtained by the random forest method, so as to specify the direction and magnitude of the correlation between the independent variable and the target variable to obtain the contribution of the independent variables.

64 variables are selected as pre-selected variables in a model from the raw data (Table 2), including 44 categorical variables and 20 continuous variables.

Table 2. 64 pre-selected variables in a model

Height	Type I Superficial muscle digital examination
Urinary tract infection	Type II Superficial muscle digital examination
Constipation	Vaginal relaxation
Fecal incontinence	Urge urinary incontinence
Hemorrhoid	Mixed urinary incontinence
Cardiac pacemaker	Leukoplakia vulvae
Cardiac metallic implants	Stress urinary incontinence
Smoking condition	Posterior vaginal wall prolapse
Hypertension	Uterine/fornix prolapse
Diabetes	If diastasis recti occurs
Chronic cough	Pre-Resting phase-score
Lumbar disease	Fast-twitch (Type II muscle fiber) phase-score
Rehabilitation and treatment condition	Slow-twitch (Type I muscle fiber) phase-score cannot be null
Education	Post-Resting phase-score
Mother or sister with Urinary incontinence	Total score
Mother or sister with Pelvic organ prolapse	Vaginal muscle strength
Current Weight	Vaginal relaxation degree
Out-patient clinic	Width of abdominal gap
Number of Induced labour	Abdominal engagement
Urinary incontinence before pregnancy	Diastasis recti
Urinary incontinence during pregnancy	Body mass index (BMI) (kg/m^2)
Urinary incontinence after delivery	Pubic symphysis separation
Gestational diabetes	Sacrococcygeal pain
Gravidity	Pain of pubic symphysis separation
Parity	Flatfoot
Total vaginal length (TVL)	Posture
Number of abortion	Degree of stress urinary incontinence
Urinary incontinence frequently occurs or not	Use urinal pad or not
Anterior vaginal wall prolapse	Urinal pad
Mother's weight gain during pregnancy_1	Urinal volume
Type I deep muscle digital examination	FBF
Type II deep muscle digital examination	Basic metabolism

The results of data modelling for severe pelvic floor dysfunction are shown in

Figure 2, where 23 relevant variables are screened out of 64 pre-selected variables, and the rankings of correlation are as follows: education, age, current weight, total vaginal length (TVL), height, gravidity, hemorrhoids, mother or sister with urinary incontinence, constipation, number of abortions, parity, number of induced labour, fetus number, history of surgeries, gestational diabetes, rehabilitation and treatment condition, lumbar disease, urinary tract infection, mother or sister with pelvic organ prolapse, chronic cough, hypertension, diabetes, and smoking condition, wherein education, total vaginal length (TVL), and height are negatively correlated, and the rest of the variables are positively correlated. The AUC of the model reaches 0.634, indicating that the screened variables could explain 63.4% of the variable in severe pelvic floor dysfunction; f1 score of the model is 0.674 with a recall rate of 0.613 in the testing set.

Further, in this embodiment, a survey is completed by 30,000 women in five communities in Chengdu with high risk factors for pelvic floor dysfunction, and Logistic and Random Forest methods are used to obtain risk factor rankings. And a risk early warning model for anterior vaginal wall prolapse, uterine/fornix prolapse, posterior vaginal wall prolapse, stress urinary incontinence, vaginal relaxation degree, and pelvic floor muscle strength (Figure 3) is initially constructed. The Random forest model for the stress urinary incontinence has the lowest accuracy of 0.70, while the classical decision tree model for vaginal relaxation degree has the highest accuracy of 0.97.

It can be demonstrated that the accuracy of the risk early warning models varies from different PFD, but the vaginal relaxation degree can be predicted more accurately, and it can be further demonstrated that the output of the risk early warning model described in the embodiment can determine the high risk/non-high risk of a woman's PFD.

Embodiment 3: Construction of a risk early warning model for analyzing anterior vaginal prolapse

The analysis and construction of the risk early warning model is carried out according to the flow chart shown in Figure 4.

(I) Data processing:

The data from the Female pelvic floor dysfunction management platform (http://scpd.yunbaober.com:8686/doctor/login) from 1 January 2014 to the present, with parity greater than 0, are processed in three main steps:

In the first step, the "parity" is de-duplicated and cleaned, and the highest parity is 5. However, there are fewer patients with quadruplet and quintuplet births, and therefore the data of patients with triplet births are mainly selected for the analysis.

In the second step, time variables such as treatment date and last menstrual period are deleted, and the polytomous variables are processed using pseudo-code.

In the third step, variables with less than 3,000 records and invalid variables (invalid variables refer to variables with only one type of record, without information difference) are deleted.

The final result showed 151 valid variables and 12,456 valid samples.

(II) Statistical test and One-way logistic regression

Statistical tests are used to analyze whether there is a significant correlation between the independent variables and "if the anterior vaginal wall prolapse occurs" (target variable). Additionally, One-way logistic regression is used to analyze whether there is a significant correlation between the independent variables and " if the anterior vaginal wall prolapse occurs " (target variable).

Table 3 Sample table of statistical test information

Variable type	**ALL**	**Group 0**	**Group 1**	**Statistics**	**P_value**	**sample_n**
Urinary tract infection	677(5.435%)	172(4.36%)	505(5.933%)	12.681	<0.001	12456
Constipation	2704(21.708%)	856(21.698%)	1848(21.713%)	0	0.996	12456
Fecal incontinence	62(0.498%)	9(0.228%)	53(0.623%)	7.696	0.006	12456
Chronic cough	228(1.83%)	64(1.622%)	164(1.927%)	1.228	0.268	12456
Pre-resting phase_ abdominal muscle_mean value	0.89(0.64)	0.89(0.63)	0.9(0.65)	0.692	0.405	8219
slow-twitch phase_ abdominal muscle_mean value	0.89(0.64)	0.89(0.63)	0.9(0.65)	0.692	0.405	8219
Post-resting phase_ abdominal muscle_mean value	0.89(0.64)	0.89(0.63)	0.9(0.65)	0.692	0.405	8219
Fatigue of Type II muscle fiber	-0.33(0.64)	-0.34(0.66)	-0.33(0.63)	0.038	0.845	3452
If a vacuum extraction is used during childbirth _2	82.0(2.468%)	18.0(2.095%)	64.0(2.598%)	0.477	0.49	3322
If obstetric forceps is used during childbirth _2	87.0(2.619%)	19.0(2.212%)	68.0(2.761%)	0.553	0.457	3322
If laceration occurs during childbirth_2	883.0(26.58%)	117.0(13.62%)	766.0(31.1%)	98.821	<0.001	3322
Body Mass Index (BMI) (kg/m^2)	22.35(2.49)	22.3(2.49)	22.36(2.49)	0.274	0.601	3143
Baby birth weight_2	3.29(0.43)	3.29(0.41)	3.3(0.44)	0.057	0.811	3030

Note: sample_n refers to the sample size of the corresponding variable without missing information (the same below), and the correlation between the corresponding variable and " if the anterior vaginal wall prolapse occurs " should be analyzed with this sample size, and a P_value less than 0.05 indicates a significant correlation, and otherwise, there is no significant correlation.

Table 4 Logistic regression results information table

Variable type	**Uni OR**	**Lower Bound**	**Upper Bound**	**Uni P Value**	**sample_n**
Urinary tract infection	1.384	1.159	1.652	0	12456
constipation	1.001	0.913	1.097	0.985	12456
Type II Superficial muscle digital examination	0.961	0.907	1.017	0.168	12228
Type II Deep muscle digital examination	0.957	0.904	1.013	0.129	12227
Total vaginal length (TVL)	1.042	0.96	1.13	0.327	12149
Lowest point of posterior vaginal wall (Bp)	5.509	4.884	6.213	0	12145
Height	1.008	1	1.016	0.062	12068
Pubic symphysis separation	432177149	0	inf	0.999	3974
Flatfoot	0.992	0.13	7.577	0.994	3837
If laceration occurs during childbirth_2	2.863	2.314	3.542	0	3322
Body Mass Index (BMI) (kg/m^2)	1.011	0.978	1.045	0.532	3143
Baby birth weight_2	1.03	0.853	1.243	0.76	3030

One-way Logistic regression models are established by applying each variable with “if the anterior vaginal wall prolapse occurs”. Uni P Value less than 0.05 indicates significant correlation, while Uni OR greater than 1 indicates a positive correlation between the variable and " if the anterior wall prolapse occurs ", otherwise, the variable is negatively correlated.

The statistical test screens 84 variables that are significantly correlated to " if the anterior vaginal wall prolapse occurs ", and the logistic regression screens 78 variables that are significantly correlated to " if the anterior vaginal wall prolapse occurs " (all of them are included in the variables screened by the statistical test). As the calculation methods of the statistics are different, the results may vary, in order not to reduce the information, 84 variables are selected as the base variables for the subsequent analyses.

The variable completeness information of the obtained 84 variables is viewed.

Table 5 Table of significant variable missing information

Variable type	Missing_Count	Missing_Percentage
Urinary tract infection	0	0
Fecal incontinence	0	0
Constipation	0	0
Total score	3788	0.304110469
Fast-twitch phase_pelvic floor_maximum value	3891	0.312379576
Pre-resting phase_ pelvic floor_mean value	3921	0.314788054
Slow-twitch phase_pelvic floor_mean value	3921	0.314788054
Post-resting phase_ pelvic floor_mean value	3921	0.314788054
Fast-twitch (Type II muscle fiber) phase_score	3983	0.319765575
Post-resting phase_exam grading (pelvic floor/abdominal muscle) (variability)	4723	0.379174695
Width of abdominal gap	6552	0.526011561
Urinary incontinence frequently occurs or not	8070	0.647880539
Diastasis recti	8216	0.659601798
Type I muscle fiber strength	8996	0.722222222
If laceration occurs during childbirth_2	9134	0.73330122
Mother’s weight gain during pregnancy_2	9461	0.759553629

The bolded variables in the table are missing a lot, and if small sample datasets are screened based on the criteria that these variables are not missing, a lot of information of other variables will be lost. Therefore, the subsequent analysis splits the data into two data sets for analysis, wherein the variables with a missing rate of less than 40% and the target variables form dataset 1, and the variables with a missing rate of more than 40% and the target variables form dataset 2. The correlation between each variable and " if the anterior vaginal wall prolapse occurs " are analyzed respectively in the datasets 1 and 2.

(III) RFE for screening variable in the model

Dataset 1 (containing 12,456 samples)

SVM is selected as the base model, and the AUC is used as the evaluation indicator to apply the RFE method combined with 5-fold cross-validation for variable screening, and finally 20 variables are selected, and the corresponding curves of the RFE variable data and the AUC values are obtained, as shown in Figure 5. When the number of variables is 20, the AUC value of model reached the highest, these 20 variables have the closest correlation with the target variables, so these 20 variables can be used to construct the model to predict the target variables.

Dataset 2 (deleting samples with more than 2/3 missing information, remaining 4,447 samples)

SVM is selected as the base model, and the AUC is used as the evaluation indicator to apply the RFE method combined with the 5-fold cross-validation for variable screening, and finally five variables are selected, and the corresponding curves of the RFE variable data and the AUC values are obtained, as shown in Figure 6.

(IV) Modelling analysis

Applying 1000 bootstrap replications, the data are sampled to divide the training set and testing set, and SHAP plots related to the variables are obtained, see Table 6 and Table 7 below and Figures 7 to 10.

Dataset 1 results are:

Table 6 Results of Testing set

model	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**	**F1**

LR	0.91(0.902, 0.919)	0.805(0.794, 0.819)	0.758(0.742, 0.775)	0.909(0.892, 0.925)	0.842(0.832, 0.854)
SVM	0.913(0.903, 0.921)	0.804(0.791, 0.816)	0.74(0.722, 0.757)	0.945(0.932, 0.957)	0.839(0.826, 0.85)
GNB	0.887(0.877, 0.898)	0.734(0.721, 0.748)	0.64(0.623, 0.659)	0.939(0.926, 0.952)	0.767(0.754, 0.781)
RF	0.914(0.904, 0.922)	0.835(0.823, 0.846)	0.899(0.887, 0.91)	0.694(0.669, 0.721)	0.882(0.873, 0.89)
GBM	0.922(0.914, 0.93)	0.849(0.837, 0.86)	0.872(0.858, 0.885)	0.798(0.775, 0.821)	0.888(0.879, 0.897)
ADA	0.92(0.91, 0.927)	0.842(0.829, 0.853)	0.871(0.858, 0.884)	0.777(0.754, 0.8)	0.883(0.874, 0.892)
MLP	0.915(0.907, 0.923)	0.833(0.821, 0.844)	0.88(0.868, 0.891)	0.731(0.705, 0.755)	0.879(0.869, 0.888)

Dataset 2 results are:

Table 7 Results of Testing set

models	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**	**F1**
LR	0.669(0.632, 0.702)	0.494(0.468, 0.522)	0.425(0.398, 0.454)	0.85(0.8, 0.894)	0.585(0.558, 0.614)
SVM	0.726(0.695, 0.758)	0.475(0.448, 0.503)	0.383(0.353, 0.41)	0.955(0.924, 0.981)	0.55(0.519, 0.578)
GNB	0.715(0.685, 0.747)	0.788(0.765, 0.81)	0.901(0.884, 0.918)	0.196(0.143, 0.251)	0.877(0.862, 0.891)
RF	0.731(0.701, 0.76)	0.839(0.819, 0.858)	1.0(1.0, 1.0)	0.0(0.0, 0.0)	0.912(0.9, 0.924)
GBM	0.72(0.687, 0.754)	0.836(0.817, 0.855)	0.996(0.993, 1.0)	0.009(0.0, 0.023)	0.911(0.899, 0.922)
ADA	0.73(0.697, 0.761)	0.839(0.819, 0.857)	1.0(1.0, 1.0)	0.0(0.0, 0.0)	0.912(0.9, 0.923)
MLP	0.717(0.685, 0.748)	0.839(0.819, 0.857)	1.0(1.0, 1.0)	0.0(0.0, 0.0)	0.912(0.9, 0.923)

Combined with Table 6 described-above, GBM has the best effect, and combined with Table 7 described-above, SVM has the best effect.

As can be seen from Figure 8, grade I posterior vaginal wall prolapse has the most significant correlation with the target variable and is positively correlated, i.e., the posterior vaginal wall prolapse occurs, the more likely to have anterior vaginal wall prolapse. Secondary to the grade I posterior vaginal wall prolapse is the lowest point of posterior vaginal wall (Bp), which is also positively correlated, i.e., the greater the Bp is, the more likely to have prolapse.

The foregoing is only a preferred embodiment of the present disclosure, and is not a limitation of the present disclosure in any form. And any simple modification or equivalent change to the above embodiments based on the technical substance of the present disclosure will fall within the protection scope of the present disclosure.

Claim

1. A method of constructing a risk early warning model for female pelvic floor dysfunction, comprising steps of:

S1: obtaining risk factors for a risk population in a region;

S2: screening correlated variables from the risk factors and constructing a risk early warning model; and

S3: training the risk early warning model to obtain an optimized model.

2. The method according to claim 1, wherein in S1, the risk population in a region is a susceptible population in a selected area.

3. The method according to claim 1, wherein in S1, the risk factors comprise external-risk factors and internal-risk factors of pelvic floor, wherein the external-risk factors include age, BMI, diabetes, physical labour, chronic cough, constipation, and smoking; and the internal-risk factors include gravidity or parity, delivery mode, pelvic floor muscle strength, and levator ani muscle injury.

4. The method according to claim 1, wherein in S2, the step of screening the correlated variables from the risk factors and constructing the risk early warning model comprises the steps of:

a. screening pre-selected variables of the model by using Logistic regression method according to the risk factors;

b. screening correlated variables of the model by a support vector machine-recursive feature elimination method according to the pre-selected variables of the model; and

c. constructing a risk early warning model by Machine learning method according to the correlated variables screened.

5. The method according to claim 4, wherein in step a, in the Logistic regression method, a formula is:

$$\mathrm{logit(P)} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_\mathrm{p} \mathrm{X}_\mathrm{p}$$

wherein, P denotes to a conditional probability of a positive result, βdenotes to a partial regression coefficient, X denotes to an independent variable, β_0 denotes to a constant term, "1, 2…p" denotes to the number of the partial regression coefficient and independent variable.

6. The method according to claim 4, wherein in step b, in the support vector machine-recursive feature elimination method, the formula is:

$$w = \sum_k a_k y_k x_k$$

wherein, w denotes to a feature weight, a denotes to a feature parameter, y denotes to a category label, x denotes to a training sample, k denotes to the number.

7. The method according to claim 4, wherein in step c, the Machine learning method comprises Logistic regression method, a support vector machine method, Gaussian Naive Bayes, Random forest algorithm, Gradient boosting machine, AdaBoost ensemble learning model and multilayer perception.

8. The method according to claim 4, wherein in step c, outputs of the risk early warning model comprise a prediction accuracy AUC, a recall rate and a f1-score.

9. A risk early warning model for female pelvic floor dysfunction constructed by using the method according to any one of claims 1 to 8.

10. A computer system, comprising a memory, a processor, and a computer program storing and running on the processor, wherein the method of constructing the model according to any one of claims 1 to 8 is performed when the processor executes the computer program.