Research on the diagnosis and treatment of the dominant diseases of traditional Chinese medicine based on machine learning

yuqi tang
Hospital of Chengdu University of Traditional Chinese Medicine https://orcid.org/0000-0002-4227-1699
Dongdong Yang (412409710@qq.com)
Zechen Li
School of Automation,Chongqing University
Yu Fang
Hospital of Chengdu University of Traditional Chinese Medicine
Shanshan Gao
Hospital of Chengdu University of Traditional Chinese Medicine

Research

Keywords: TCM, Insomnia, Machine learning, Diagnosis, Association rules, Cluster analysis, Random forest

Posted Date: July 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-42369/v1

Version of Record: A version of this preprint was published on January 6th, 2021. See the published version at https://doi.org/10.1186/s13020-020-00409-8.

Research on the diagnosis and treatment of the dominant diseases of traditional Chinese medicine based on machine learning

Yuqi Tang[1], Dongdong Yang[1*], Zechen Li[2],Yu Fang[1],Shanshan Gao[1]

Corresponding author: Dongdong Yang　Email：412409710@qq.com

（1.Department of neurology，Hospital of Chengdu University of Traditional Chinese Medicine, Chengdu, 610072 CN;

2.School of Automation, Chongqing University, Chongqing,400044 CN）

Abstract

Background: Insomnia as one of the dominant diseases of traditional Chinese medicine (TCM) has been extensively studied in recent years. To explore the novel approaches of research on TCM diagnosis and treatment, this paper presents a strategy for the research of insomnia based on machine learning.

Methods: First of all, 654 insomnia cases have been collected from an experienced doctor of TCM as sample data. Secondly, in the light of the characteristics of TCM diagnosis and treatment, the contents of research samples have been divided into four parts: the basic information, the four diagnostic methods, the treatment based on syndrome differentiation and the main prescription. And then, these four parts have been analyzed by three analysis methods, including frequency analysis, association rules and hierarchical cluster analysis. Finally, a comprehensive study of the whole four parts has been conducted by random forest.

Results: Researches of the above four parts revealed some essential connections. Simultaneously, based on the algorithm model established by the random forest, the accuracy of predicting the main prescription by the combination of the four diagnostic methods and the treatment based on syndrome differentiation was 0.85. Furthermore, having been extracted features through applying the random forest, the syndrome differentiation of five zang-organs was proven to be the most significant parameter of the TCM diagnosis and treatment.

Conclusions: The results indicate that the machine learning methods are worthy of being adopted to study the dominant diseases of TCM for exploring the crucial rules of the diagnosis and treatment.

Keywords

TCM; Insomnia; Machine learning; Diagnosis; Association rules; Cluster analysis; Random forest.

Background

The application of TCM can be traced back to thousands of years[1]. In spite of the fact that TCM is still regarded as the complementary and alternative therapy in the field of modern medicine, it can hardly be ignored that TCM has attracted widespread attention in recent years due to its unique personalized treatment scheme and the outstanding treatment effect on some dominant diseases[2-3]. Insomnia is one of the dominant diseases of TCM. It has been proven that TCM has been successfully applied to the treatment of insomnia in the medical field[4-5]. Compared with the western medicine in the treatment of insomnia, the advantages of TCM treatment are the personalization of diagnosis and treatment ideas, the non-dependence of treatment drugs and the diversity of treatment schemes, etc. Unlike the diagnosis and treatment of the western medicine, which is based on rigorous scientific trials, most of TCM diagnoses are relied on the experience of doctors to get comprehensive and personalized treatment strategies. Consequently, TCM is considered as an empirical medicine as well. Nonetheless, it should be noted that a set of core theories of TCM have been established since the beginning of the TCM development. Subsequently, the core theories of TCM have been developed into the TCM prescription, acupuncture, meridians and other theories[6]. Moreover, in the long-term clinical practice, with the constant deepening of the understanding of the basic theories of TCM, the diagnosis and treatment ideas of TCM have been promoted tremendously, and the diagnosis and treatment standards have achieved an innovation as well[7]. Diagnosis and treatment ideas and treatment strategies are the critical points of the clinical practice. Meanwhile, the medical record data are the embodiment of diagnosis and treatment ideas, thus worth exploring. The medical record of TCM is composed of four parts, including the basic

information, the four diagnoses of TCM, the treatment based on syndrome differentiation and the main prescription. The diagnosis and treatment of TCM is a whole from the information collection (including basic information and four diagnoses) to the treatment based on syndrome differentiation, and then to the establishment of the main prescription. The whole diagnosis and treatment process is not only logical, but also indivisible. In the past decades, many efforts have been done to study this process, whereas most researches have only focused on one part of this process. Zhang S et al. applied the data mining technology to explore the drug rules of pulmonary fibrosis based on TCM medical records[8]. Yu XW et al. analyzed the dose data of TCM prescriptions by optimizing the traditional Cheng-Church double clustering algorithm (CC)[9]. Liu y et al. adopted the data mining method to verify the TCM syndrome patterns of PSCI[10]. These researches have shown some opinions on the diagnosis and treatment process of TCM to some extent. However，their research methods have violated the core principle of integration and personalization of the TCM diagnosis and treatment, resulting that their conclusions can hardly be applied in clinical practice[11]. Therefore, for the sake of reliability and comprehensiveness of the research method adopted in the present paper, the research is carried out logically according to the sequence of TCM diagnosis and treatment, and the whole will be discussed at last.

In recent years, the rapid development of data analysis and artificial intelligence has provided a novel research direction for the improvement of the clinical diagnosis and treatment technology. At present, the methods of data mining and machine learning have been widely used in the field of TCM[12].

In the present paper, the medical record data of insomnia are selected as the research samples.

Based on the medical record data, the research method of diagnosis and treatment of insomnia of TCM is emphatically discussed by applying machine learning methods. Specifically, the above-mentioned four parts in the process of TCM diagnosis and treatment are analyzed separately by three analysis methods, including frequency analysis, association rules and hierarchical cluster analysis. And then, a thorough analysis of the whole four parts is conducted using random forest. Considering that the data used in each analysis step have unique characteristics, different analysis schemes are established for different parts of the data.

1 Data and methods

1.1 Sample data

The sample data are obtained from the Sichuan Provincial Hospital of TCM under the confidentiality agreement and the authority approval. According to the Guidelines for the diagnosis and treatment of insomnia in China(2017)[13] and the International Classification of Sleep Disorders(ICSD2)(2005)[14], the inclusion criteria are set as follows: the medical record data should contain one or more symptoms below: ①Sleep latency (SL) is prolonged and more than 30 minutes; ②Having difficulty in sleep maintenance, mainly manifested by easy and early to wake up;③The quality of sleep is decreased, and the patient can hardly get into deep sleep and have multiple dreams; ④Insufficient sleep duration (less than 6.5 hours); ⑤With daytime symptoms, including fatigue, emotional problems, memory and attention decline, daytime sleepiness and work initiative decline, etc. The exclusion criteria are set as: ①The missing of the medical record data is so severe that it is unable to meet the research requirements; ② The patients have other serious organic diseases that

may cause insomnia.

In our preliminary work, 1577 outpatient data (from 2016 to 2020) are collected and screened according to the above inclusion and exclusion criteria. Finally, only 654 outpatient data are selected as the research samples. Since the selection and analysis of medical record data of TCM have a high demand for expertise, three professional doctors of TCM are selected to analyze, code and classify the medical record data information of research samples manually. Meanwhile, the workload is equally assigned to the three doctors, and the cross-validation is implemented after all work has been completed, so as to eliminate the impact of subjectivity and artificial errors on the final data. And then, the sample database is established. Simultaneously, according to the TCM diagnosis and treatment ideas, the contents of the sample data are divide into four parts: the basic information, the four diagnostic treatment, the treatment based on syndrome differentiation and the main prescription. Each part contains several data, and the specific data processing steps will be described later. In the light of the characteristics of the data, the machine learning methods, including frequency analysis, association rules and hierarchical clustering analysis, are adopted to process and mine the data. Finally, the data of the TCM diagnosis and treatment ideas from the four diagnoses, the treatment based on syndrome differentiation and the main prescription are integrally discussed by employing the random forest algorithm. The specific research strategy designed in this paper is illustrated in **Figure 1.Flowchart of the research strategy designed in this paper**.

In the process of coding and classification, the Guidelines for the diagnosis and treatment of insomnia in China(2017)[13] and the International Classification of Sleep Disorders(ICSD2)(2005)[14] are regarded as the basis to ensure the objectivity and

comprehensiveness of the data. In the meantime, based on the combination of TCM insomnia related symptoms and the syndrome differentiation, a complete code comparative table is shown in **Table 1**. Three doctors of TCM are required to complete their work in strict accordance with the code comparative table.

1.2 Data processing and machine learning

1.2.1 Data preprocessing

Data preprocessing consists of data alignment, missing value processing and data format conversion, etc. It is worth mentioning that the medical record information is extracted strictly according to the coding table, and there are a extremely small number of incomplete cases in the actual medical records. The incomplete items are represented by null values in the process of data set making. To eliminate the impact of the null value on the research and ensure that the follow-up research process can be carried out smoothly, the substitute values are selected to fill the null values of the record data. The substitute values include the course of disease and sleep duration, etc., and these values are filled with their mean value. The substitute values are specified in **Table 2**.

The processed data set are import into Python. The data samples are quantified by programming, and then analyzed by applying the following machine learning methods.

1.2.2 Frequency analysis

Frequency is also known as "time". The total data are divided into groups according to the preset standards, and then the number of individuals in each group is counted. The relative frequency is the ratio of the frequency of each group to the total number of data.

1.2.3 Association rules

A frequently-used method to study the relationship rules among data is to apply the association rules of Apriori algorithm[15]. Generally, three indicators, including confidence, support and lift, can be used to evaluate an association rule. Support is defined as the proportion of the data in the item set to the data in the data set, thus measuring the frequency of a set appearing in the original data. For instance, if two sets in the data set are X and Y respectively, then：

$$Support(X \to Y) = P(X \mid Y) \qquad (1)$$

where X|Y represents the union of X and Y.

Confidence is defined for an association rule. The confidence of X→Y can be expressed as follows:

$$Confidence = p\{x \mid y\} \Big/ P\{X\} \qquad (2)$$

Lift can reflect the correlation between X and Y in association rules. As expressed in the following function, the lift is defined as the proportion of the probability of the data set containing both X and Y to the probability of the data set only containing Y.

$$Lift(X \to Y) = P(Y \mid X) \Big/ P(Y) \qquad (3)$$

The higher the lift is (lift＞1), the higher the positive correlation is，and vice versa. The lift equal to 1 indicates that there is no correlation.

1.2.4 Cluster analysis

At present, the cluster analysis is extensively used in the medical field[16]. In general, the cluster analysis can be classified into two categories, one is hierarchical clustering algorithm and the other is agglomerative clustering algorithm. In the Euclidean space, using hierarchical clustering algorithm to analyze small-scale data sets can achieve optimal results. Its basic principle is to establish a hierarchical clustering tree by calculating the similarity among different categories of data points and adopting the bottom-up aggregation strategy. Each sample set in the data sets is regarded as a cluster, and then the clusters with close distance are merged step by step to achieve the expected number of clusters.

Assuming that there are clusters C_i and C_j, the function can be described as follows:

$$D_{aug}(C_i, C_j) = \frac{1}{|C_i \| C_j|} \sum_{x \in C_i} \sum_{z \in C_j} dist(x, z) \qquad (4)$$

where the average distance $D_{aug}(C_i, C_j)$ is determined by all samples of the two clusters.

1.2.5 Random forest

The random forest algorithm derived from ensemble learning method is composed of multiple decision trees. The random forest is an extension of the classification tree and the regression tree. These trees can be used to model the response variables through recursive partition and predict the final results jointly[17]. The random forest algorithm is commonly employed in data classification and regression[18]. At present, there are three mainstream decision tree algorithms, including ID3, C4.5 and CART. In the present paper, the most widely used algorithm, CART, is selected to build random forest algorithm model. The main function of this algorithm is described below.

Suppose that there is a training data set D with k classes in total. The Gini index of set D can be expressed as follows:

$$Gini(D) = \sum_k \frac{|C_k|}{D}(1 - \frac{|C_k|}{D}) = 1 - \sum_k (\frac{|C_k|}{|D|})^2 \quad (5)$$

where C_k represents the sample subset of class k. The $|C_k|$ and $|D|$ represent the size of C_k and D respectively.

In CART algorithm，assuming that feature A is used to segment the data. If feature A is a discrete feature, set D can be divided into subset D1 and subset D2 according to one possible value a of A, as shown below.

$$D_1 = \{D \mid A = a\}; D_2\{D \mid A \neq a\} \quad (6)$$

Consequently, the *Gini*(*D*,*A*) of set D under the condition of known feature A can be obtained by combining the above functions. The Gini index is theoretically similar to entropy, as described below.

$$Gini(D, A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \quad (7)$$

Similar to the principle of entropy, the greater the value of *Gini*(*D*,*A*) is, the greater the sample uncertainty is. Taking this into account, the value of *Gini*(*D*,*A*) should be as small as possible when selecting the feature A.

2 Results and discussion

2.1 Basic information

The basic information mainly consists of the ID, name, clinic time, age and gender of patients. Since the clinic time is not taken as a factor in the screening criteria during the data screening stage, the statistical results may deviate from the actual situation. The ID and name of patients have no impact on the diagnosis and treatment process. As a consequence, the focus of this section is age and

gender of patients. Considering that the categories of age and gender data are relatively few, we choose frequency analysis for the data processing. According to the box-plot of age distribution of patients (shown in **Figure 2**.**Box-plot of the age distribution of patients**), it can be seen that the average age of patients is 47, the mean square deviation of age is 11. Moreover, the maximum age and minimum age are 79 and 14 respectively.

Figure 3.Pie chart of the gender distribution of patients depicts the gender distribution of patients. As illustrated in this figure, the proportion of female patients is significantly higher than that of male patients.

2.2 Four diagnostic methods

Four diagnostic methods include inspection (observation), auscultation and olfaction (listening and smelling), interrogation (inquirying or questioning) and palpation (pulse examination). Basically, it is a process of collecting medical history information for doctors of TCM[19]. “Inspection” refers to the observation of patients' external performance, such as tongue picture, expression, reaction and complexion. Moreover, “auscultation and olfaction” is the way that doctors diagnose diseases by hearing and smelling. Additionally, “interrogation” is a sort of diagnostic method for doctors to find out the occurrence, development, treatment process and past health history of diseases by talking with patients. Furthermore, “palpation” particularly refers to the method that doctors use index fingers, middle fingers and ring fingers to touch the special position of radial artery of patients to check the pathological changes of patients. In this paper, the four diagnostic methods are further classified on the basis of the characteristics of the medical record data of insomnia research samples (shown in **Figure 4.Classification of four diagnostic methods**).

Based on the smallest unit of classification, the method of association rules is applied to study in this section. Considering that the basic information is also a part of TCM interrogation and may have an effect on the diagnosis and treatment process of the diseases, the basic information is included in the four diagnostic parts for discussion as well. Taking into account that there are too many null values in some of the smallest classification units, we attempt to use two methods to analyze the association rules for the combination of the smallest units (the combination items are in the brackets below), so as to minimize the impact of the null values on the research results. The results are listed in **Table 3** and **Table 4**.

Method 1: four diagnostic methods of TCM: (tongue proper, tongue color and tongue coating), (sleep duration, sleep status, course of insomnia, concomitant symptoms and emotion), pulse, age and gender.The results are summarized in **Table 3**.

Method 2: (tongue proper, tongue color and tongue coating), pulse (sleep duration, sleep status and course of insomnia), emotion, (concomitant symptoms, others), age and gender. The results are summarized in **Table 4**.

It can be concluded from the above tables that most of the results are dominantly related to gender and age, while there is no significant association among the four diagnoses. According to the analysis of the actual clinical experience, the above results have no remarkable guiding significance for clinical practice. Nevertheless, two innovative research directions can be found based on these results. On the one hand, this research can be explored deeply through expanding the sample size and using other methods to find the internal association of the four diagnoses. On the other hand, there is no obvious external association among the data, but these data have the statistical significance. These

data can be used for the epidemiological study of TCM on condition that the sample size is large enough.

2.3 Treatment based on syndrome differentiation

Originating from the philosophical culture, the treatment based on syndrome differentiation is the core of the TCM theories and gradually develops into a complex theoretical framework, including the yin and yang theory, five elements, eight principles, the Qi and blood theory，the organs theory and the meridian system[20].

The treatment based on syndrome differentiation is a comprehensive analysis by doctors in the process of diagnosis and treatment of TCM, and its judgment criteria are derived from the objective medical record information including the four diagnoses. In essence, the treatment based on syndrome differentiation is a high generalization of the causes of different patients according to the principle of personalized treatment. In the light of the logic of TCM syndrome differentiation, the treatment based on syndrome differentiation can be divided into three parts, including the eight principal syndrome differentiation, the organs syndrome differentiation and the meridian syndrome differentiation. Further, these three parts can be separated into several items. The previous studies have either skipped this process directly or determined the syndrome differentiation category only based on the experience description of doctors, which were too empirical. Based on the characteristics of insomnia in TCM, this paper focuses on four significant syndrome differentiation points, namely the syndrome differentiation of asthenia and sthenia, the syndrome differentiation of cold and heat, the syndrome differentiation of organs and pathogenic factors. The medical record data are extracted by three professional doctors of TCM, and then classified and coded according to the above four significant syndrome differentiation points. It is worth mentioning that the organs

syndrome differentiation includes heart, liver, spleen, lung, kidney, gall bladder, stomach, small intestine, large intestine, bladder and the triple burner; the syndrome differentiation of asthenia and sthenia consists of asthenia syndrome and sthenia syndrome; the syndrome differentiation of cold and heat is composed of cold syndrome and heat syndrome; the pathogenic factors include phlegm, fire, blood stasis and asthenia. The above-mentioned 19 syndrome differentiation factors constitute the section of treatment based on syndrome differentiation of the insomnia sample data research in this paper. To ensure the objectivity of each syndrome differentiation factor, the three TCM doctors are supposed to collect at least two or more kinds of medical record information in the classification and coding stage of medical record data for determining one syndrome differentiation factor. For instance, the medical information "wiry pulse" and "irritability" can infer that the syndrome differentiation factor of organs is liver; the medical information "thin pulse" and "tiredness" can imply that the factor of asthenia and sthenia syndrome differentiation is asthenia syndrome; the medical information "red tongue" combined with "tidal fever" and "rapid pulse " indicates that the factor of cold and heat syndrome differentiation is heat syndrome; the medical information "slippery pulse" combined with "yellow tongue" and "greasy tongue coating" means that the pathogenic factors is phlegm.

Despite that each syndrome differentiation factor in each medical record is relatively independent, there is a strong correlation among the factors. Therefore, it is reasonable to select association rules for the analysis. Through a process of trial and error, the confidence is finally adjusted to 0.7, and the results are summarized in **Table 5**.

As can be seen from the above table, besides the associations that can be obtained from the basic theories, such as the associations between fire and sthenia syndrome, fire and heat syndrome,

there are more new-found associations. For example, the complex syndrome of heart, liver, spleen, asthenia and sthenia → the heat syndrome, the fire stasis syndrome → the heart, liver. The following conclusions can be drawn by analyzing the treatment based on syndrome differentiation with association rules. On the one hand, the results can reveal the syndrome differentiation thoughts of TCM doctors. On the other hand, after applying the above methods to classify the contents of treatment based on syndrome differentiation, the results can reflect the priority direction of syndrome differentiation of insomnia to a certain extent, thus having guiding significance for clinical practice. In the further study, more research methods can be adopted to verify the dominant diseases of TCM and explore new syndrome differentiation rules.

2.4 Main prescription

Basically, the treatment strategy is composed of acupuncture, moxibustion, scraping therapy and TCM prescription, etc. In the present paper, the TCM prescription is the research focus of this section. TCM prescription is the embodiment of clinical practice of TCM. Choosing the appropriate combinations of Chinese medicine under the guidance of the treatment based on syndrome differentiation not only reflects the typical thoughts of TCM, but also conforms to the treatment method of drug combination therapy[21]. Having completed the process from the four diagnoses to the treatment based on syndrome differentiation, the doctors should determine the main prescription. And then, on the basis of the main prescription, the doctors should adjust the prescription properly according to the actual situation of patients. Finally, the treatment prescription can be obtained. Thus, the determination of the main prescription is particularly significant. The main prescription can not only prove the personalized treatment advantages of TCM, but also reflect the most core treatment method in the clinical practice of TCM. The previous studies have achieved some success; however,

there are two deficiencies in their researches. First, the previous researches have mainly focused on the frequency of herb use and interrelation of the herbs. Second, there are few previous researches concerned about the components of the main prescription[22-23]. Taking the above deficiencies into account, the less use herbs are removed from the statistics of the herb use frequency, thus reducing the impact on the research of the main prescription in this paper. **Table 6** presents the correspondence between the processed data codes and herbs.

For the sake of reducing calculation amount and the increasing the code execution efficiency, all the herbs are replaced with codes, and then the codes are entered into the database.

The hierarchical clustering algorithm is employed to analyze the small sample data set in the European space, thus obtaining a satisfactory result. According to the characteristics that the main prescription is composed of a wide variety of herbs, the hierarchical clustering algorithm is applied to explore the potential classification rules in the data samples of TCM. The results of the analysis of the main prescription using hierarchical clustering analysis are shown in **Figure 5.Results of the analysis of the main prescription using hierarchical clustering analysis**.

The main prescriptions of the corresponding serial number are presented in **Table 7**, and the repeated herb combinations in all main prescriptions are shown in **Table 8**.

The above conclusions indicate that the desired results can be achieved by adopting the hierarchical clustering algorithm to analyze the main prescriptions. The rapid acquisition of the main prescription of TCM is beneficial for the study of the combination rules of TCM, but also lays a solid foundation for the overall study of the diagnosis and treatment of the dominant diseases of TCM. In order to facilitate the further discussion on the overall diagnosis and treatment idea, we code the main prescriptions of TCM and enter the codes into the database.

2.5 Diagnosis and treatment idea

In the discussion of the aforementioned four parts, the four parts of TCM diagnosis and treatment ideas are studied successively, so as to reveal the internal relationship and related research methods of each part. This section discusses the four parts as a whole. In accordance with the research process designed in the previous section(in Figure 1), the random forest algorithm is adopted to establish the model. Simultaneously, the data sets collected from four diagnoses, treatment based on syndrome differentiation, and the main prescriptions of TCM are put into the model for cross-validation. Consequently, the corresponding accuracy can be obtained. In the meantime, for the purpose that the internal relationship of TCM diagnosis and treatment ideas can be explored deeper, this section is divided into two processes for further discussion. These two process are illustrated in **Figure 6.Flowchart of the diagnosis and treatment ideas of TCM**.

It is worth noting that five zang-organs, six fu-organs and pathogenic factors each contains several syndrome differentiation factors, which are randomly combined in the medical record sample data. In addition, in the actual outpatient service, the prescriptions made by doctors for patients commonly includes at least one main prescription. Therefore, in order to facilitate data processing, the five zang-organs combination, six fu-organs combination, pathogenic factors combination and main prescription combination are coded and loaded into the database. For the sake of presenting the accuracy more intuitively, the method of confusion matrix is carried out in this paper. The confusion matrix results are shown in **Figure 7.Confusion matrix**.

As summarized in **Table 9**, the accuracy of applying the random forest algorithm model to predict the information of treatment based on syndrome differentiation through the four diagnostic information is dramatically high. Simultaneously, the high accuracy is achieved by predicting the

main prescription through the information of the combination of the four diagnoses and the treatment based on syndrome differentiation.

In process 2 of this section, the random forest algorithm model is applied to extract the eigenvalues of all data in the data sets. Since the eigenvalues obtained by using the random forest model are too small to be studied conveniently, the eigenvalues are expanded in the form of logarithmic transformation to facilitate the observation. The transformed eigenvalues are shown in **Figure 8.Transformed eigenvalues obtained by using random forest model**.

As illustrated in Figure 8, the most significant parameter affecting the judgment results is the syndrome differentiation of five zang-organs , followed by sleep status, pulse conditions, the syndrome differentiation of asthenia and sthenia and the syndrome differentiation of six fu-organs. Meanwhile, emotion status, pathogenic factors and tongue picture (including tongue proper, tongue color and tongue coating) also have a tremendous effect on the judgment results. Nevertheless, sleeping duration, insomnia course, syndrome differentiation of cold and heat, and other items except the tongue picture in the inspection and the auscultation and olfaction have less influence on the selection of the final main prescription. As can be seen from the above results, doctors take the sleep status, pulse conditions and tongue picture as the most critical indicators when they are obtaining the four diagnoses information. In the meantime, the emotional status is also taken into account for understanding the basic situation of the patient's condition. Based on the the syndrome differentiation of five zang-organs, and combined with the syndrome differentiation of asthenia and sthenia and the syndrome differentiation of six fu-organs, a comprehensive analysis is conducted to obtain the final main prescription in the process of syndrome differentiation. Since sleep duration, course of

insomnia and other factors have little impact on the diagnosis and treatment process, they are only regarded as reference for the diagnosis and treatment.

It can be concluded from the above results that the random forest algorithm model can be applied to quickly and accurately verify the correctness of TCM diagnosis and treatment ideas. It is worth mentioning that only one algorithm model is used in this paper, resulting in the lack of the diversity of methods. In the further research, a wide variety of algorithm models can be introduced for comparisons, so as to further investigate the feasibility of machine learning methods in the research of TCM diagnosis and treatment.

3 Conclusions

The results indicate that the machine learning methods can be effectively applied to deeply mine and analyze the medical record data of the dominant diseases of TCM. The focus of this study is to analyze the diagnosis and treatment process of the TCM dominant diseases which includes the acquisition of the patients' condition information through using four diagnostic methods, and the flexible application of the syndrome differentiation methods to develop the treatment plan and select the main prescription. And the research strategy established in this paper can efficiently filter the unessential diagnosis and treatment information, thus helping TCM doctors to quickly and efficiently obtain valuable information and crucial rules from a substantial number of medical record data. Furthermore, since the research process, the data collection and the data analysis methods designed in this paper are highly standardized, the research strategy established in this paper can be applied to further investigate the diagnosis and treatment rules of other TCM dominant disease.

List of abbreviations

1	Traditional Chinese medicine	TCM [1]
2	Cheng-Church double clustering algorithm	CC
3	Sleep latency	SL

Declarations

Ethics approval and consent to participate

Informed consent of the study and a statement on ethics approval was waived because of the retrospective nature and the analysis used anonymous clinical data.

Consent for publication

Not applicable

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

Not applicable

Authors' contributions

YT designed the study and write the manuscript.ZL implemented the thought with software, analyzed the data.SG and YF dedicated in experiment results analysis and manuscript revision.DY participated into analysis implementation and organized discussion of the results.All authors read and approved the final manuscript.

Acknowledgements

Thanks to Mengyu Chen for her professional writing assistance.

References

1、Gao H, Wang Z, Li Y, Qian Z. Overview of the quality standard research of traditional Chinese medicine. Front Med. 2011;5(2):195-202.

2、Yan D, Liu J, Wang AT, Yang ZR, Yue SJ, Feng XZ. Exploring Research Ideas of Mechanism of Dominant Diseases in Traditional Chinese Medicine Based on Evidence-Based Medicine.Zhongguo Zhong Yao Za Zhi. .2018;43(13):2633-2638.

3、Chen YB, Tong XF, Ren J, Yu CQ, Cui YL. Current Research Trends in Traditional Chinese Medicine Formula: A Bibliometric Review from 2000 to 2016. Evid Based Complement Alternat Med. 2019;2019:3961395.

4、Zhang H, Liu P, Wu X, Zhang Y, Cong D. Effectiveness of Chinese herbal medicine for patients with primary insomnia:A PRISMA-compliant meta-analysis. Medicine (Baltimore). 2019;98(24):e15967.

5、Li F, Xu B, Wang P, Liu L. Traditional Chinese medicine non-pharmaceutical therapies for chronic adult insomnia: A Bayesian network meta-analysis protocol. Medicine (Baltimore). 2019;98(46):e17754.

6、Li Z, Xu C. The fundamental theory of traditional Chinese medicine and the consideration in its research strategy. Front Med. 2011;5(2):208–211.

7、Wang J, Guo Y, Li GL. Current Status of Standardization of Traditional Chinese Medicine in China. Evid Based Complement Alternat Med. 2016;2016:9123103.

8、Zhang S, Wu H, Liu J, Gu H, Li X, Zhang T. Medication regularity of pulmonary fibrosis treatment by contemporary traditional Chinese medicine experts based on data mining. J Thorac Dis. 2018;10(3):1775–1787.

9、Yu XW, Gong QY, Hu KF, Mao WJ, Zhang WM. Research on Ratio of Dosage of Drugs in Traditional Chinese Prescriptions by Data Mining. Stud Health Technol Inform. 2017;245:653–656.

10、Liu Y, Liu D, Zhang Y, et al. Markov Clustering Analysis-Based Validation for Traditional Chinese Medicine Syndrome Patterns of Poststroke Cognitive Impairment. J Altern Complement Med. 2019; 25(11):1140–1148.

11、Zhou X, Li Y, Peng Y, et al. Clinical phenotype network: the underlying mechanism for personalized diagnosis and treatment of traditional Chinese medicine. Front Med. 2014;8(3):337–346.

12、Zhao C, Li GZ, Wang C, Niu J. Advances in Patient Classification for Traditional Chinese Medicine: A Machine Learning Perspective. Evid Based Complement Alternat Med. 2015;2015:376716.

13、Han F,Tang XD,Zhang B.The Guidelines for the diagnosis and treatment of insomnia in China.Natl Med J China.2017,97(24):1844-1856.

14、American Academy of Sleep Medicine . International Classification of Sleep Disorders: Diagnostic and Coding Classification of Sleep Disorders697 Manual. 2nd ed. Westchester: American Academy of Sleep Medicine; 2005.

15、Somek M, Hercigonja-Szekeres M. Decision Support Systems in Health Care - Velocity of Apriori Algorithm. Stud Health Technol Inform. 2017;244:53–57.

16、Xu R, Wunsch DC 2nd. Clustering algorithms in biomedical research: a review. IEEE Rev Biomed Eng. 2010;3:120-154.

17、Jones FC, Plewes R, Murison L, et al. Random forests as cumulative effects models: A case study of lakes and rivers in Muskoka, Canada. J Env iron Manage. 2017;201:407–424.

18、Sun J, Yu H, Zhong G, Dong J, Zhang S, Yu H. Random Shapley Forests: Cooperative Game-Based Random Forests With Consistency. IEEE Trans Cybern. 2020;10.1109/TCYB.2020.2972956.

19、Kang H, Zhao Y, Li C, et al. Integrating clinical indexes into four-diagnostic information contributes to the Traditional Chinese

Medicine (TCM) syndrome diagnosis of chronic hepatitis B. Sci Rep. 2015;5:9395.

20、Ma Y, Zhou K, Fan J, Sun S. Traditional Chinese medicine: potential approaches from modern dynamical complexity theories. Front Med. 2016;10(1):28–32.

21、Jin ZL, Hu JX, Jin HW, Zhang LR, Liu ZM. Analysis of Traditional Chinese Medicine Prescriptions Based on Support Vector Machine and Analytic Hierarchy Process.Zhongguo Zhong Yao Za Zhi. 2018;43(13):2817-2823.

22、Zhang S, Wu H, Liu J, Gu H, Li X, Zhang T. Medication regularity of pulmonary fibrosis treatment by contemporary traditional Chinese medicine experts based on data mining. J Thorac Dis.2018;10(3):1775-1787.

23、Leem J, Jung W, Kim Y, Kim B, Kim K. Exploring the combination and modular characteristics of herbs for alopecia treatment in traditional Chinese medicine: an association rule mining and network analysis study. BMC Complement Altern Med. 2018;18(1):204.

The figure legends

Figure 5.Results of the analysis of the main prescription using hierarchical clustering analysis.(Page 16)

The codes in Figure 5 correspond to the codes in Table 6. The line represents the the main prescription, and the small circle represents the corresponding herb. Frequency of Category 1 is 573, frequency of Category 2 is 312, frequency of Category 3 is 577.

Figure 6.Flowchart of the diagnosis and treatment ideas of TCM. **(Page 17)**

Process 1: the information of the treatment based on syndrome differentiation is deduced from the data of four diagnoses. The information of the treatment based on syndrome differentiation includes five parts: cold and heat, asthenia and sthenia, five zang-organs, six fu-organs and pathogenic factors.

Process 2: The main prescription is deduced from the four diagnostic information and the information of the treatment based on syndrome differentiation.

Figure 7.Confusion matrix. **(Page 17)**

Process 1 is shown in Figure 7(A-E). **Process 2** is presented in Figure7(F).In the confusion matrix, the vertical coordinate is the diagnosis made by doctors in the original medical records, and the horizontal coordinate represents the predicted value made by the random forest. Taking the "cold and heat" confusion matrix (as depicted in Figure 7(A)) in process 1 as an example, the cold and heat syndrome can be derived from the data of four diagnoses. The total number of medical record

samples is 654, including 73 cases without cold and heat syndrome, 28 cases with cold syndrome, 419 cases with heat syndrome, and 134 cases with cold and heat complex syndrome. As shown in Figure 7(A), among the predicted values of the random forest model, the numbers of the cases accurately predicted by the random forest model for the above syndromes are 55, 14 , 418 and 94 respectively. In general, a total of 581 cases are accurately predicted, and the prediction accuracy is 0.89. Similarly, the information of asthenia and sthenia, five zang-organs, six fu-organs and pathogenic factors can be derived from the the data of four diagnoses, and the numbers of the cases accurately predicted are 611, 576, 562 and 557 respectively

Table 1 Code comparative table of various insomnia related symptoms (Page 6)

Item	Content	Code	Item	Content	Code	Item	Content	Code
Tongue color	Normal	0	**Insomnia course**	≤3 months	1	**Sleep duration**	Normal	0
	Pale	1		3 months-1 year	2		Pernoctation	1
	Red	2		1-3 years	3		0-1 hours	2
	Dark	3		3-5 years	4		1-2 hours	3
	Others	4		＞5 years	5		2-3 hours	4
Cold and heat	Normal	0	**Asthenia and sthenia**	Normal	0		3-4 hours	5
	Cold	1		Asthenia	1		4-5 hours	6
	Heat	2		Sthenia	2		5-6 hours	7
	Cold and heat complex	3		Asthenia and sthenia Complex	3		6-7 hours	8
	Others	4		Others	4		＞7 hours	9
Five zang-organs	Normal	0	**Pathogenic factors**	Normal	0		Taking medicine to sleep	A
	Heart	1		Phlegm	1	**Emotional status**	Normal	0
	Liver	2		Fire	2		Anxious	1
	Spleen	3		Blood stasis	3		Fear	2
	Lung	4		Asthenia	4		Nervous	3
	Kidney	5		Others	5		Restlessness	4
Tongue proper	Normal	0	**Sleeping status**	Normal	0		Timidity	5
	Enlarged tongue	1		Difficult to fall asleep	1		Irritability	6
	Thin tongue	2		Dysphylaxia	2		Depressed	7
	Teeth print on tongue	3		Festless sleep	3		Flusteredness	8
	Cleft tongue	4		Hard to fall asleep again after waking up	4		Others	9
	The vessels of sublingual The tongue with ecchymosis	5		Dreaminess	5		Normal	0
	Others	6		Others	6		Headache	1
Six fu-organs	Normal	0	**Others**	Normal	0		Dizziness	2
	Stomach	1		Sallow complexion	1		Lethargic	3
	Gallbladder	2		Flushed cheeks	2		Aversion to cold	4
	Large intestine	3		Redness of the eyes	3		Aversion to heat	5
	Small intestine	4		Dark lip	4	**Concomitant symptoms**	Aversion to cold or heat Irregular	6
	Bladder	5		Halitosis	5		Tidal fever	7
	Sanjiao	6		Eyes bright	6		Night sweat	8
				Others	7		Snoring	9
Pulse conditions	Normal	0	**The tongue coating**	Normal	0		Nocturia	A
	Thin	1		Yellow	1		Fatigue	B
	Wiry	2		Thin	2		Dry mouth	C
	Slippery	3		Slimy	3		Bitter taste in	D
	Rapid	4		White	4		Abnormal stool and urine	E
	Deep	5		Scanty	5		Others	F
	Floating	6		Thick	6			
	Others	7		Dye	7			
				Others	8			

Table 2 Comparative table of the null values and the substitute values (Page 7)

Item	Substitute value
Six fu-organs combination	Normal
Tongue proper	Normal
Pulse conditions	Normal
Emotional status	Normal
Five zang-organs	Heart
Insomnia course	1-3 years
Asthenia and sthenia	Asthenia and sthenia complex
Cold and heat	Cold and heat complex
Tongue color	Others
Pathogenic factors	Others

Table 3 Summary of the results using Method 1 (Page 12)

Association	Confidence	Support	Lift
39-year old→Female	1.0	0.02	1.37
59-year old→Female	1.0	0.02	1.37
56-year old→Female	1.0	0.02	1.37
29-year old→Female	1.0	0.02	1.37
（Tongue proper：normal/Tongue color：pale/Tongue coating：thin、yellow/Pulse conditions：thin、wiry）→Female	1.0	0.02	1.37
（Tongue proper：normal/Tongue color：pale/Tongue coating：thin、yellow）→Female	1.0	0.04	1.31
（Tongue proper：teeth print on tongue/Tongue color：red、dark/Tongue coating：thin、yellow）→Female	1.0	0.03	1.3
（Tongue proper：normal/Tongue color：pale、dark/Tongue coating：thin、yellow）→Female	1.0	0.03	1.29
（Sleep duration:normal/Sleeping status:Difficult to fall asleep/Insomnia cours ＞5 years)→Female	0.9	0.03	1.29
（Tongue proper：teeth print on tongue/Tongue color：red/Tongue coating：thin、yellow/Pulse conditions：wiry、rapid）→Female	0.9	0.02	1.28
62-year old→Female	0.9	0.02	1.28
50-year old→Female	0.9	0.05	1.28
30-year old→Female	0.9	0.02	1.25

Table 4 Summary of the results using Method 2 (Page 12)

Association	Confidence	Support	Lift
（Sleep duration:3-4 hours /Sleeping status:Difficult to fall asleep/Insomnia course:3 months-1 year)→47-year old	1.0	0.01	14.22
（Sleep duration:3-4 hours /Sleeping status:Difficult to fall asleep/Insomnia course:3 months-1 year)→Male	1.0	0.01	3.72
57-year old→Female	1.0	0.01	1.37
38-year old→Female	1.0	0.01	1.37
39-year old→Female	1.0	0.02	1.37
34-year old→Female	1.0	0.01	1.37
59-year old→Female	1.0	0.02	1.37
56-year old→Female	1.0	0.02	1.37
29-year old→Female	1.0	0.02	1.37
Pulse conditions：thin、wiry、rapid、deep→Female	1.0	0.01	1.37
Pulse conditions：wiry、rapid、deep→Female	1.0	0.01	1.37
（Tongue proper：normal/Tongue color：pale/Tongue coating：thin、white）→Female	1.0	0.01	1.37

Table 5 Summary of the results analyzing the syndrome differentiation factors by adopting association rules (Page 14)

Association	Confidence	Support	Lift
Pathogenic factors：fire→Asthenia and sthenia：sthenia	1	0.08	3.86
Pathogenic factors：fire→Cold and heat：heat	1	0.08	1.56
Pathogenic factors：fire/Cold and heat：heat→Asthenia and sthenia：sthenia	1	0.08	3.86
Asthenia and sthenia：sthenia/Pathogenic factors：fire→Cold and heat：heat	1	0.08	1.56
Pathogenic factors：fire→Asthenia and sthenia：sthenia/Cold and heat：heat	1	0.08	1.45
Pathogenic factors：fire、blood stasis→Asthenia and sthenia：sthenia	0.96	0.07	3.7
Pathogenic factors：fire、blood stasis→Cold and heat：heat	0.96	0.07	1.49
Pathogenic factors：fire、blood stasis、asthenia→Asthenia and sthenia：asthenia and sthenia complex	0.92	0.08	3.22
Asthenia and sthenia：sthenia→Cold and heat：heat	0.92	0.26	1.43
Pathogenic factors：fire→Five zang-organs：heart、liver	0.89	0.08	2.8
Pathogenic factors：asthenia→Asthenia and sthenia：asthenia	0.89	0.17	1.97
Pathogenic factors：fire、blood stasis→Five zang-organs：heart、liver	0.85	0.07	2.69
Five zang-organs：heart、liver/Six fu-organs：gallbladder→Cold and heat：heat	0.85	0.13	1.33
Five zang-organs：heart、liver→Cold and heat：heat	0.85	0.32	1.32
Cold and heat：normal→Asthenia and sthenia：asthenia	0.82	0.11	1.83
Five zang-organs：heart、liver、spleen/Asthenia and sthenia：asthenia and sthenia complex→Cold and heat：heat	0.82	0.1	1.28
Five zang-organs：heart、liver、spleen/Six fu-organs：normal→Cold and heat：heat	0.8	0.08	1.25
Pathogenic factors：phlegm、asthenia→Asthenia and sthenia：asthenia	0.79	0.09	1.75

Five zang-organs：heart、spleen→Asthenia and sthenia：asthenia	0.76	0.21	1.69
Five zang-organs：heart、spleen/Six fu-organs：normal→Asthenia and sthenia：asthenia	0.75	0.09	1.67
Pathogenic factors：blood stasis、asthenia→Asthenia and sthenia：asthenia	0.73	0.2	1.63
Five zang-organs：heart、liver、spleen→Cold and heat：heat	0.73	0.23	1.14
Asthenia and sthenia：asthenia and sthenia complex→Cold and heat：heat	0.72	0.28	1.12

Table 6 Correspondence between the herbs and the codes (Page 15)

Chinese herbal medicine	Code	Chinese herbal medicine	Code
Spine date seed	1	Pinellia ternate	26
Glycyrrhiza	2	White peony root	27
Anemarrhena	3	Atractylodes Macrocephala	28
Poria cocos	4	Prepared radix rehmanniae	29
Ligusticum wallichii	5	Chinese yam	30
Caulis polygoni multiflori	6	Cornus officinalis	31
Lily	7	Cortex moutan	32
Seed of oriental arborvitae	8	Magnolia officinalis	33
Red peony root	9	Tasteless preserved soybean	34
Gentian	10	Arillus longan	35
Scutellaria	11	Astragalus	36
Gardenia	12	White hyacinth bean	37
Alisma orientale	13	Villous amomum	38
Caulis Aristolochiae Manshuriensis	14	Semen coicis	39
Plantain	15	Os draconis (longgu)	40
Angelica sinensis	16	Oyster	41
Radix rehmanniae	17	Lanceolata	42
Ginseng	18	Radix auckladiae	43
Polygala	19	Placenta hominis	44
Schisandra chinensis	20	Blighted wheat	45
Coptis chinensis	21	Leonurus japonicus	46
Bamboo shavings	22	Cinnamon	47

Citrus aurantium	23	Eucommia	48
Tangerine peel	24	Jianqu	49
Bupleurum	25		

Table 7 Correspondence between the main prescriptions and the serial numbers (Page 16)

Serial number	Main prescription
1	Spine date seed、Glycyrrhiza、Anemarrhena、Poria cocos、Ligusticum wallichii、Caulis polygoni multiflori、Lily、Seed of oriental arborvitae、Red peony root
2	Bupleurum、Pinellia ternate、Astragalus、Os draconis (longgu)、Oyster
3	Gentian、Scutellaria、Gardenia、Alisma orientale、Caulis Aristolochiae Manshuriensis、Plantain、Angelica sinensis、Radix rehmanniae
4	Angelica sinensis、White peony root、Atractylodes Macrocephala
5	Ginseng、Poria cocos、Polygala、Angelica sinensis、Schisandra chinensis、Seed of oriental arborvitae、Radix rehmanniae、Spine date seed
6	Coptis chinensis、Bamboo shavings、Citrus aurantium、Tangerine peel
7	Bamboo shavings、Citrus aurantium、Tangerine peel
8	Bupleurum、Pinellia ternate、Astragalus
9	White peony root、Atractylodes Macrocephala、Villous amomum、Ginseng、Chinese yam、Semen coicis
10	Atractylodes Macrocephala、Angelica sinensis、Arillus longan、Polygala、Astragalus
11	Prepared radix rehmanniae、Chinese yam、Cornus officinalis、Cortex moutan
12	Magnolia officinalis
13	Tasteless preserved soybean、Gardenia

Table 8 Correspondence between the repeated herb combinations and the serial numbers (Page 16)

Serial number	Combination
1	Spine date seed、Poria cocos、Seed of oriental arborvitae
2	Bupleurum、Pinellia ternate、Astragalus
3	Bamboo shavings、Citrus aurantium、Tangerine peel
4	White peony root、Atractylodes Macrocephala
5	Angelica sinensis、Radix rehmanniae
6	Angelica sinensis、Polygala

Table 9 Prediction accuracy of applying the random forest algorithm model (Page 17)

Item	Accuracy
Cold and heat	0.89
Asthenia and sthenia	0.93
Five zang-organs	0.88
Six fu-organs	0.86
Pathogenic factors	0.85
Main prescription	0.85

Figures

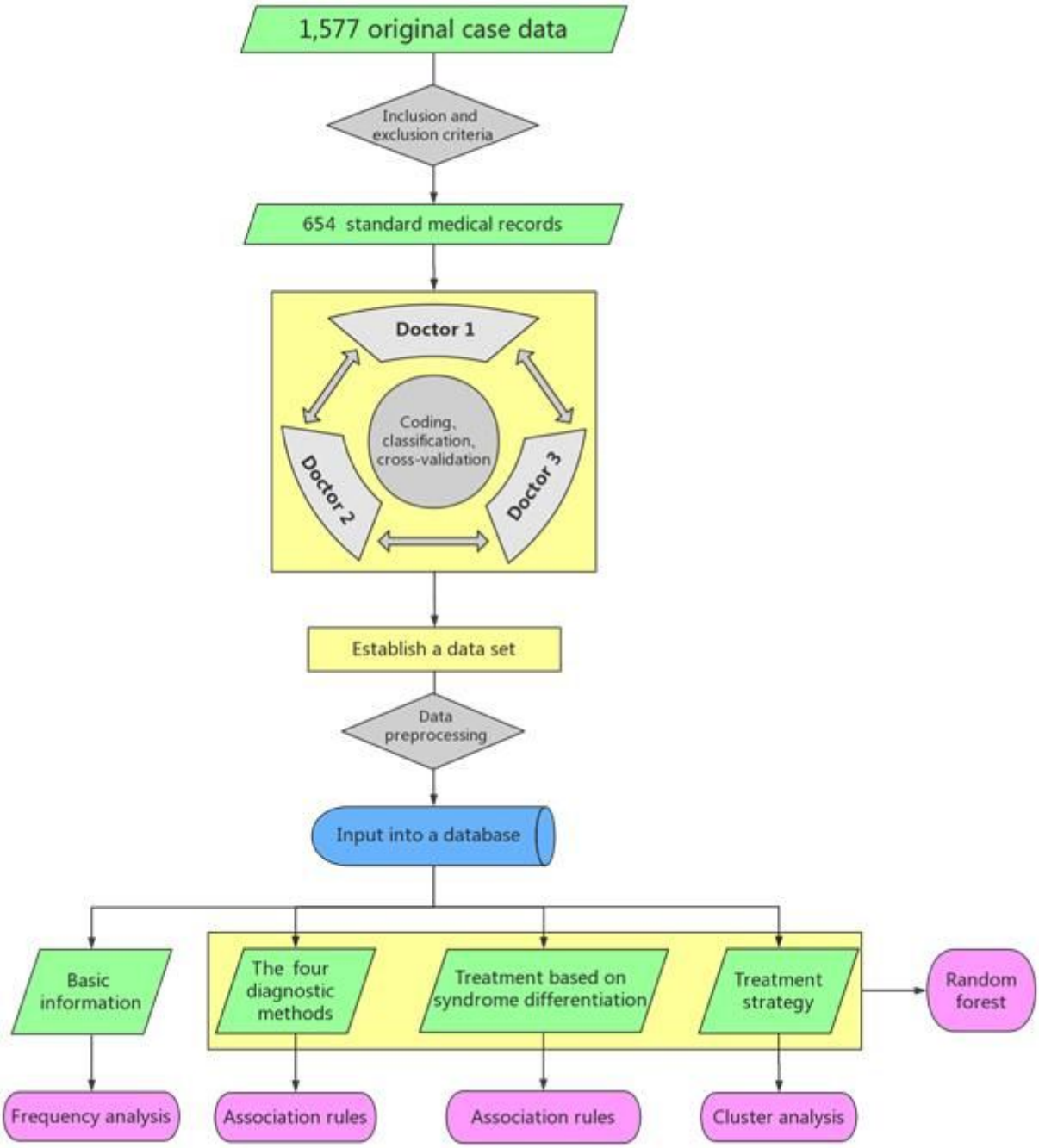

Figure 1

Flowchart of the research strategy designed in this paper.

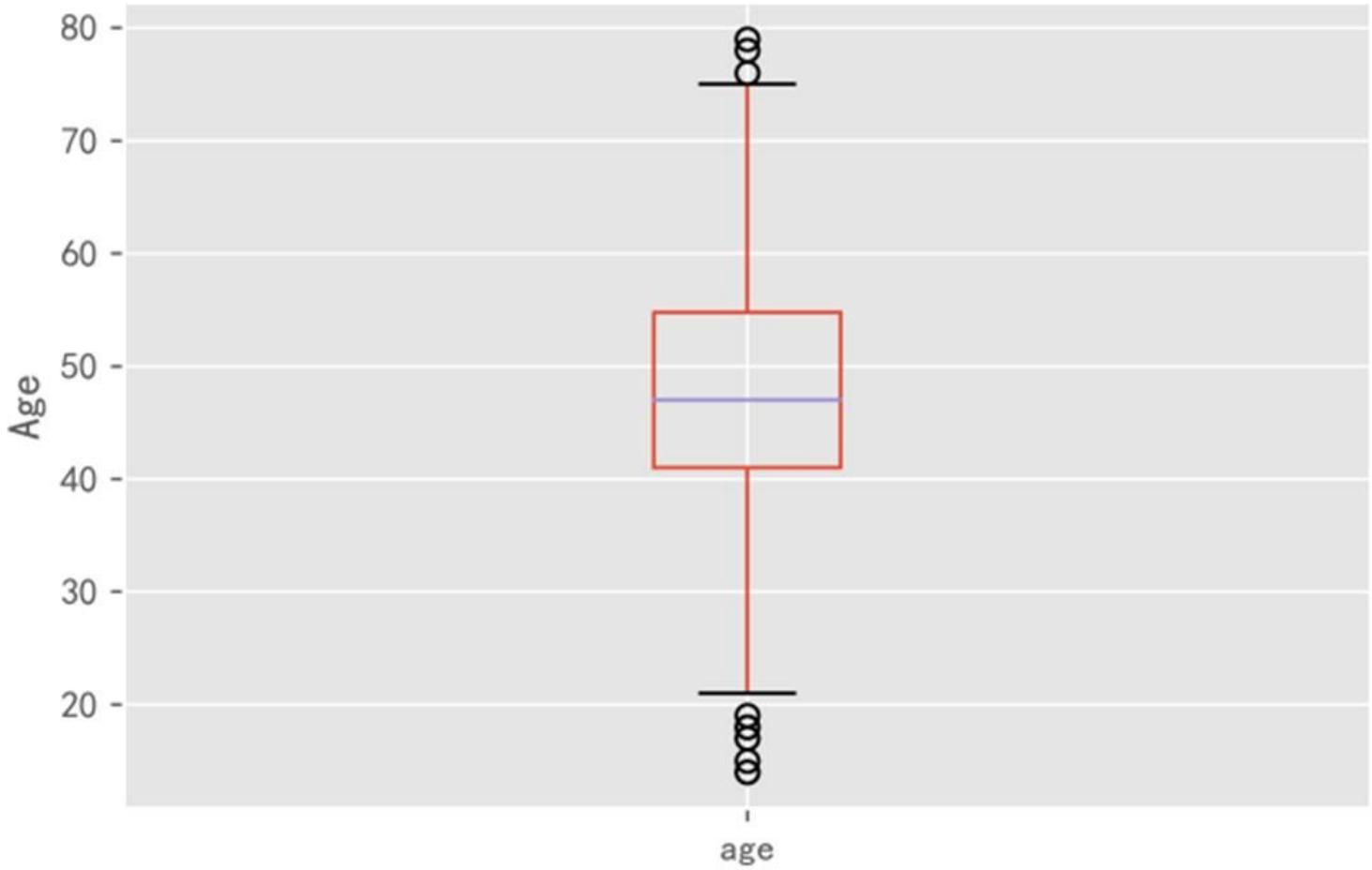

Figure 2

Box-plot of the age distribution of patients

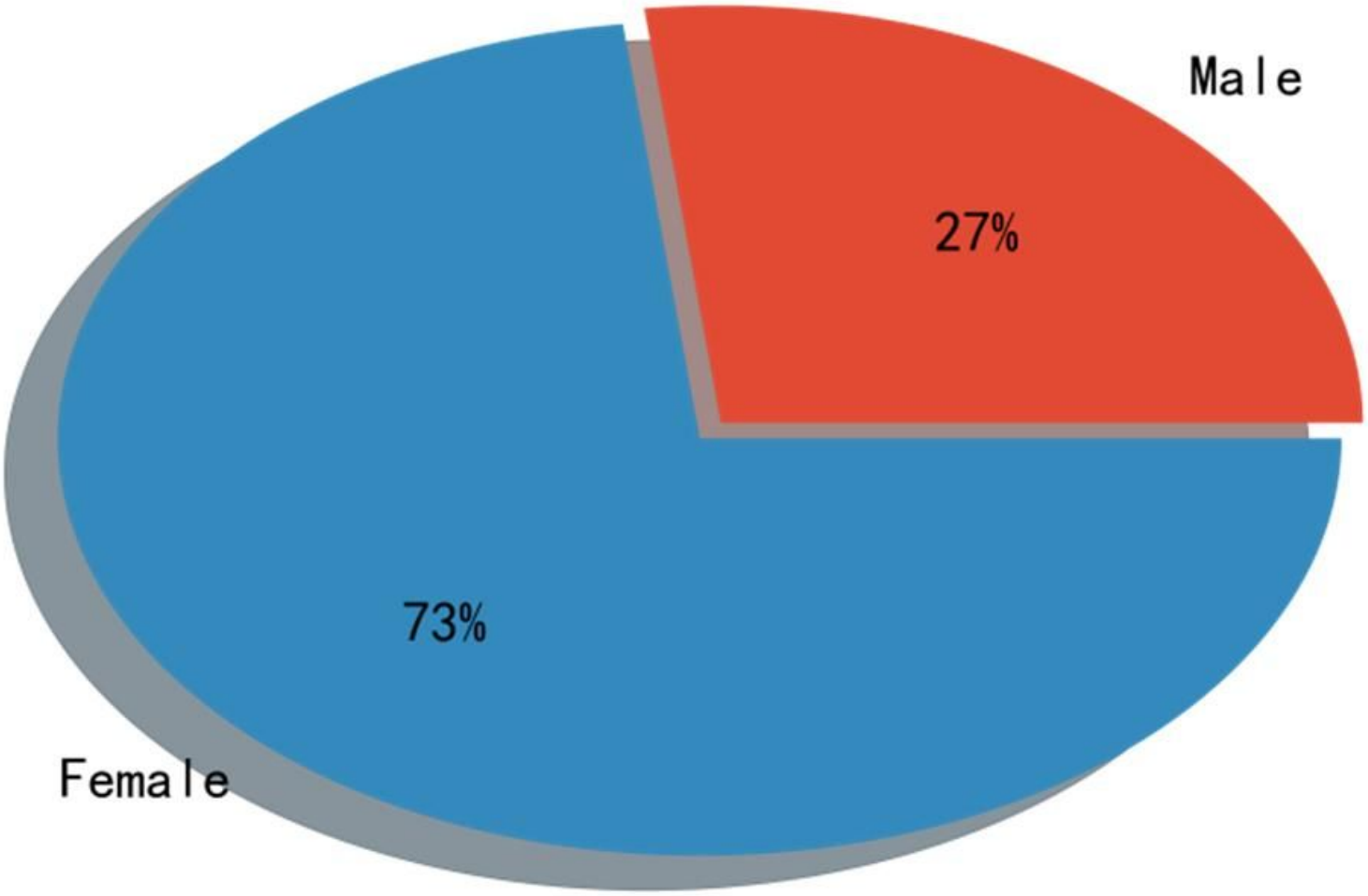

Figure 3

Pie chart of the gender distribution of patients

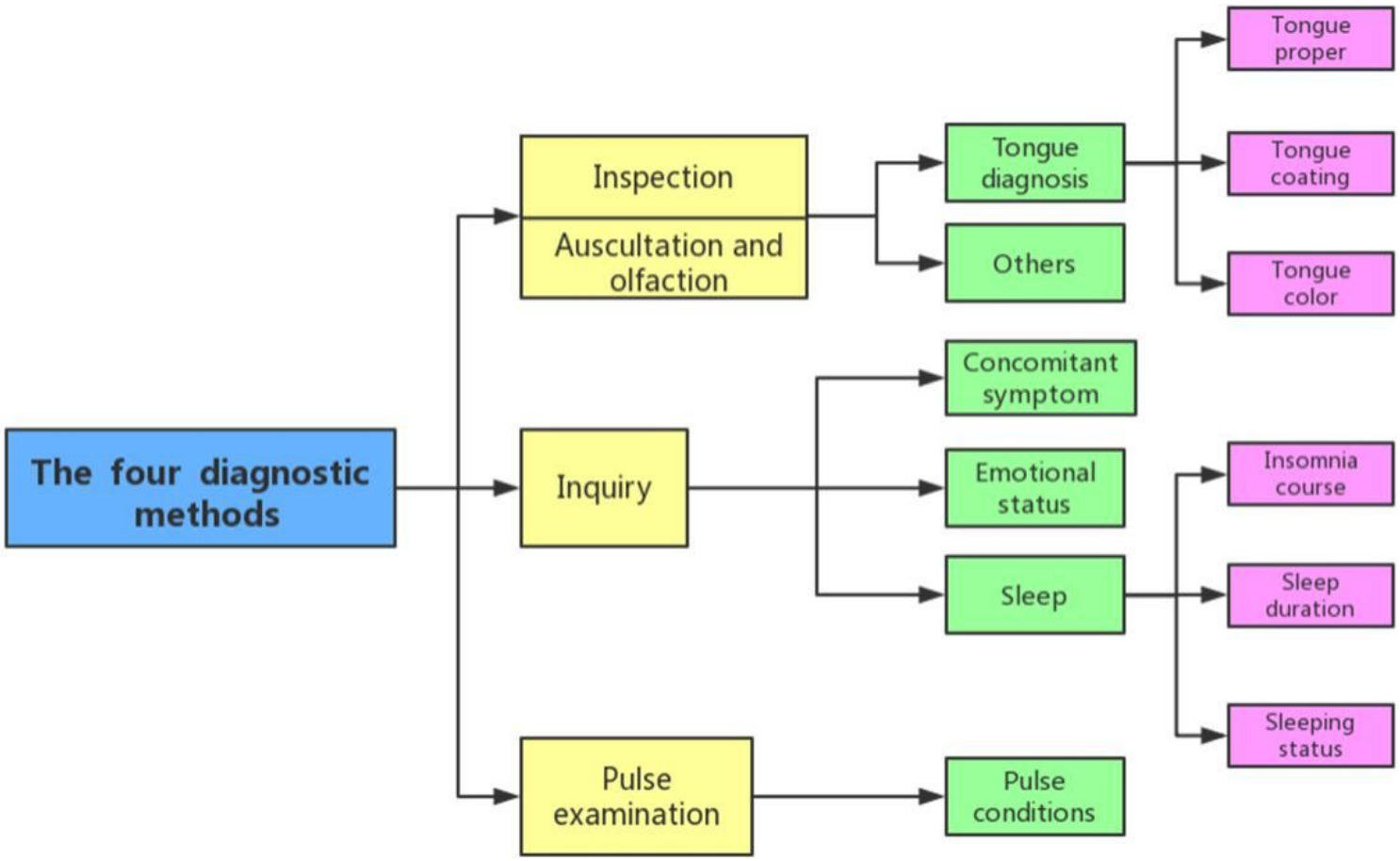

Figure 4

Classification of four diagnostic methods

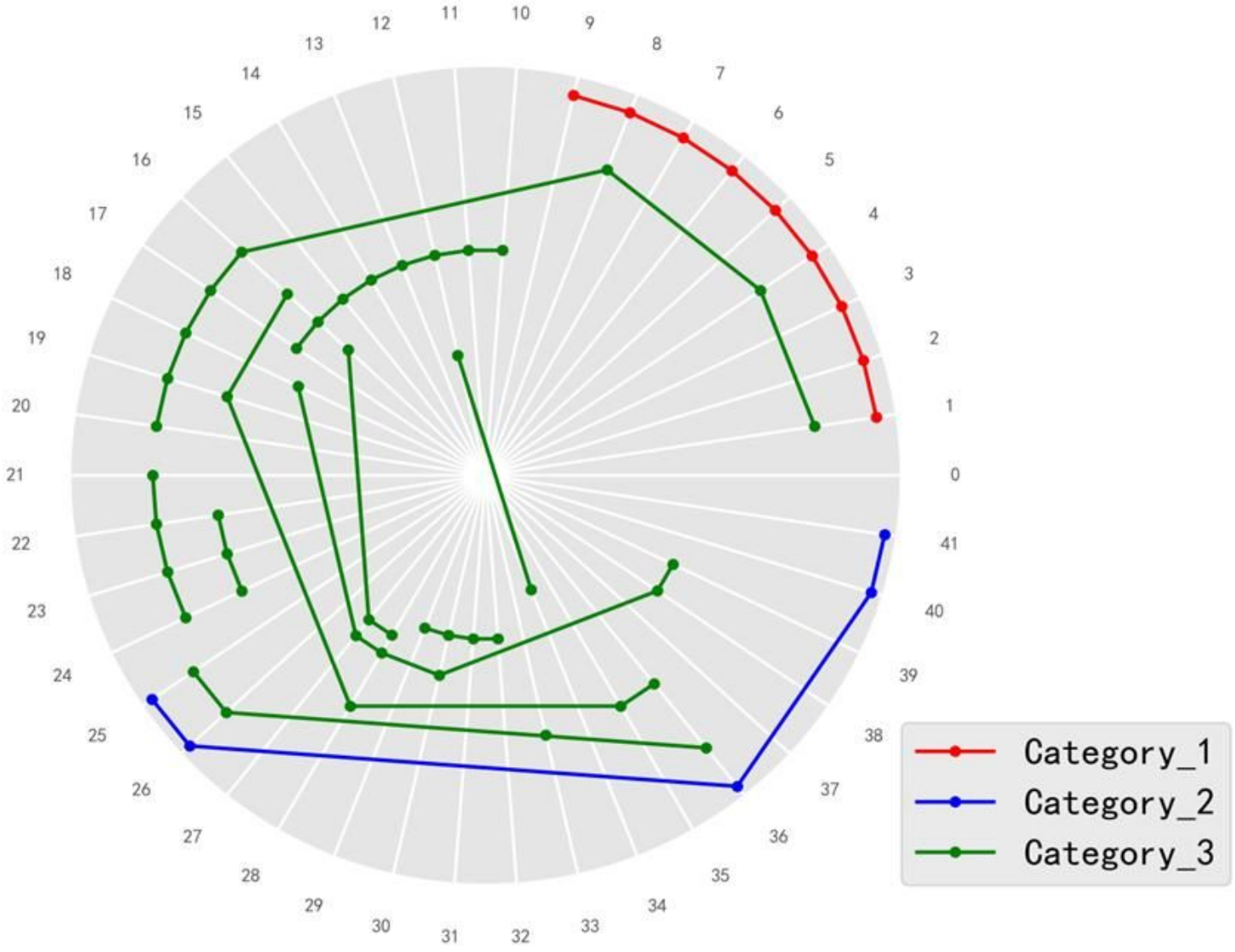

Figure 5

Results of the analysis of the main prescription using hierarchical clustering analysis.(Page 16) The codes in Figure 5 correspond to the codes in Table 6. The line represents the the main prescription, and the small circle represents the corresponding herb. Frequency of Category 1 is 573, frequency of Category 2 is 312, frequency of Category 3 is 577.

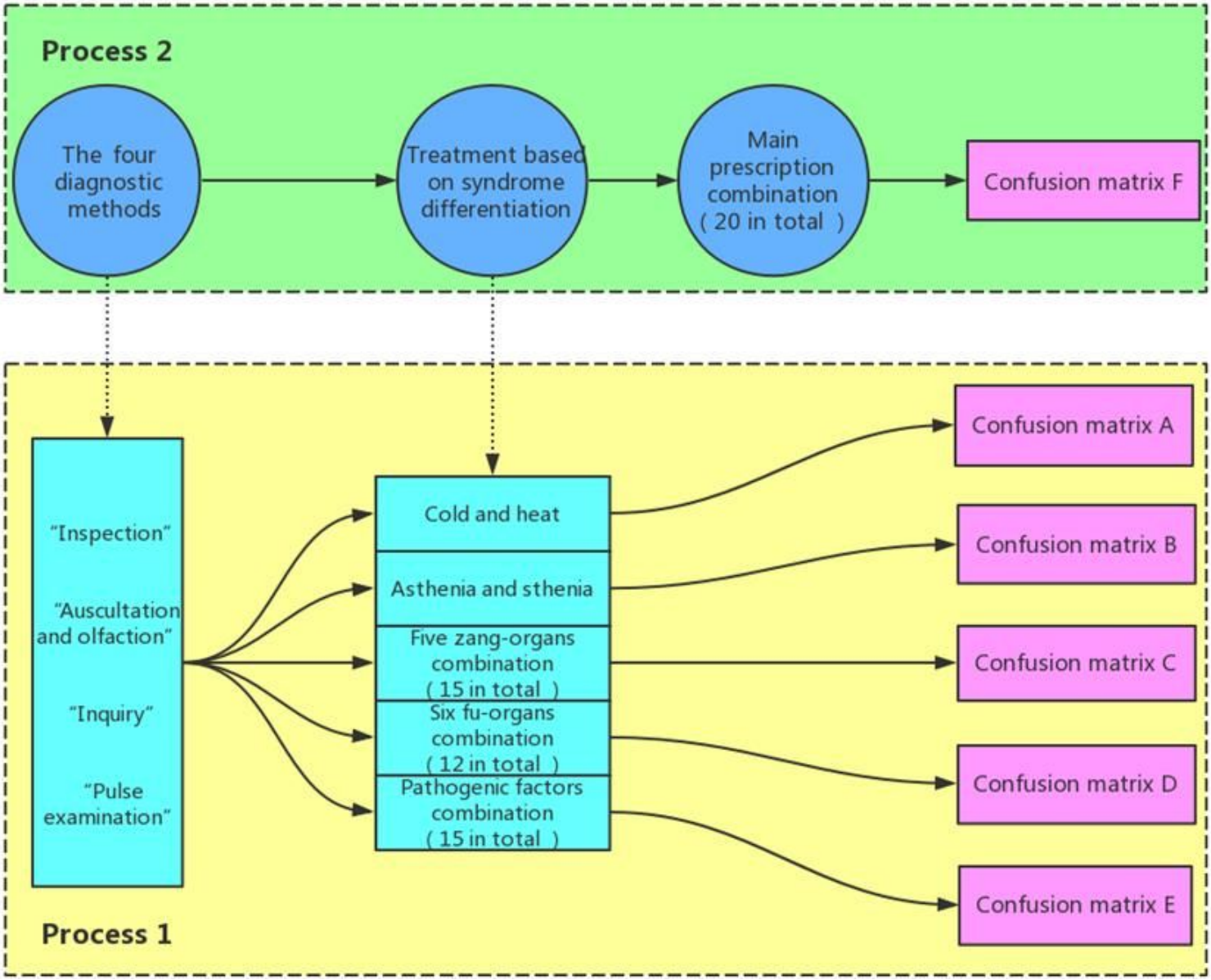

Figure 6

Flowchart of the diagnosis and treatment ideas of TCM. (Page 17) Process 1: the information of the treatment based on syndrome differentiation is deduced from the data of four diagnoses. The information of the treatment based on syndrome differentiation includes five parts: cold and heat, asthenia and sthenia, five zang-organs, six fu-organs and pathogenic factors. Process 2: The main prescription is deduced from the four diagnostic information and the information of the treatment based on syndrome differentiation.

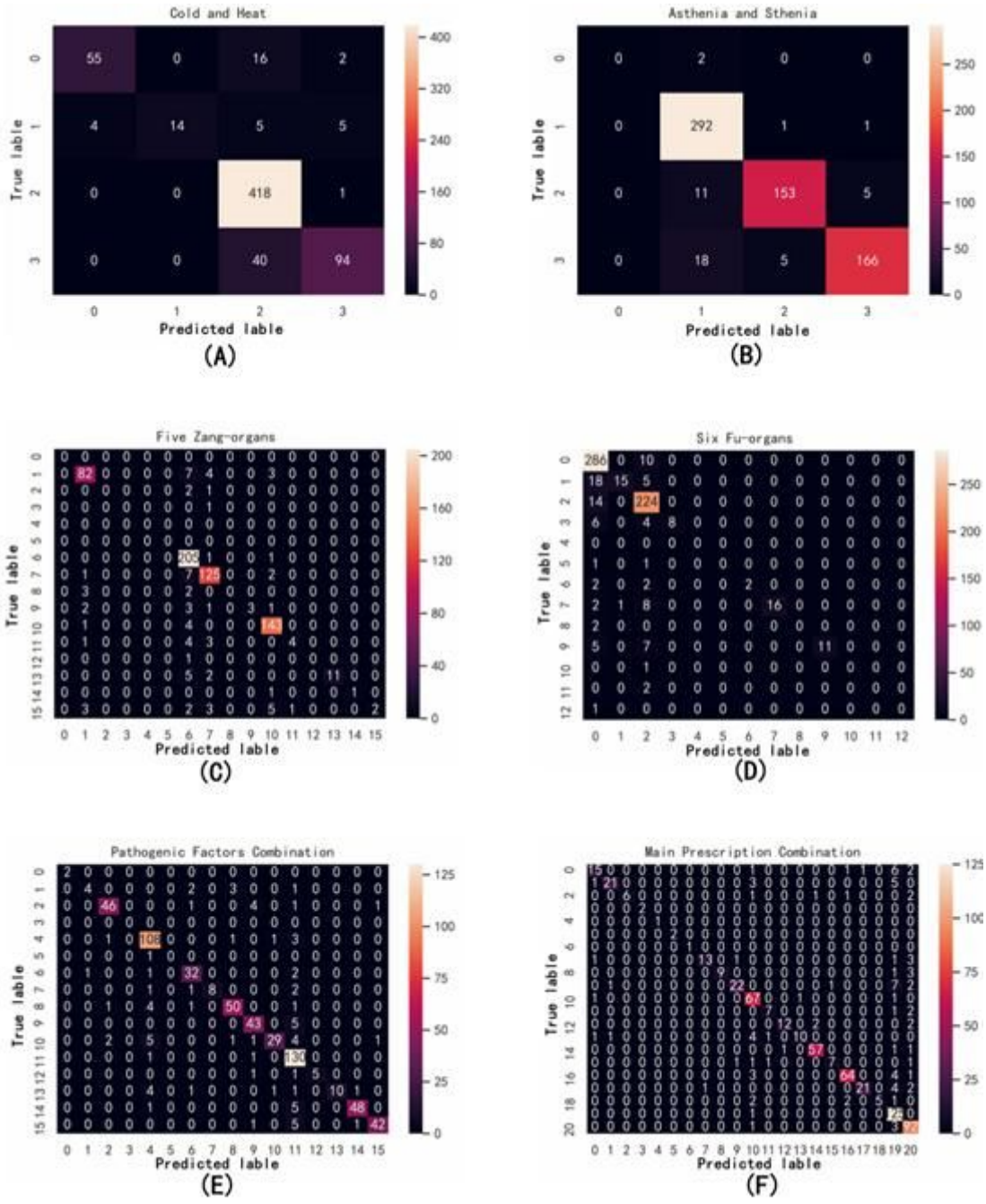

Figure 7

Confusion matrix. (Page 17) Process 1 is shown in Figure 7(A-E). Process 2 is presented in Figure7(F).In the confusion matrix, the vertical coordinate is the diagnosis made by doctors in the original medical records, and the horizontal coordinate represents the predicted value made by the random forest. Taking the "cold and heat" confusion matrix (as depicted in Figure 7(A)) in process 1 as an example, the cold and heat syndrome can be derived from the data of four diagnoses. The total number of medical record samples is 654, including 73 cases without cold and heat syndrome, 28 cases with cold syndrome, 419 cases with heat syndrome, and 134 cases with cold and heat complex syndrome. As shown in Figure 7(A), among the predicted values of the random forest model, the numbers of the cases accurately predicted by the random forest model for the above syndromes are 55, 14 , 418 and 94 respectively. In

general, a total of 581 cases are accurately predicted, and the prediction accuracy is 0.89. Similarly, the information of asthenia and sthenia, five zang-organs, six fu-organs and pathogenic factors can be derived from the the data of four diagnoses, and the numbers of the cases accurately predicted are 611, 576, 562 and 557 respectively

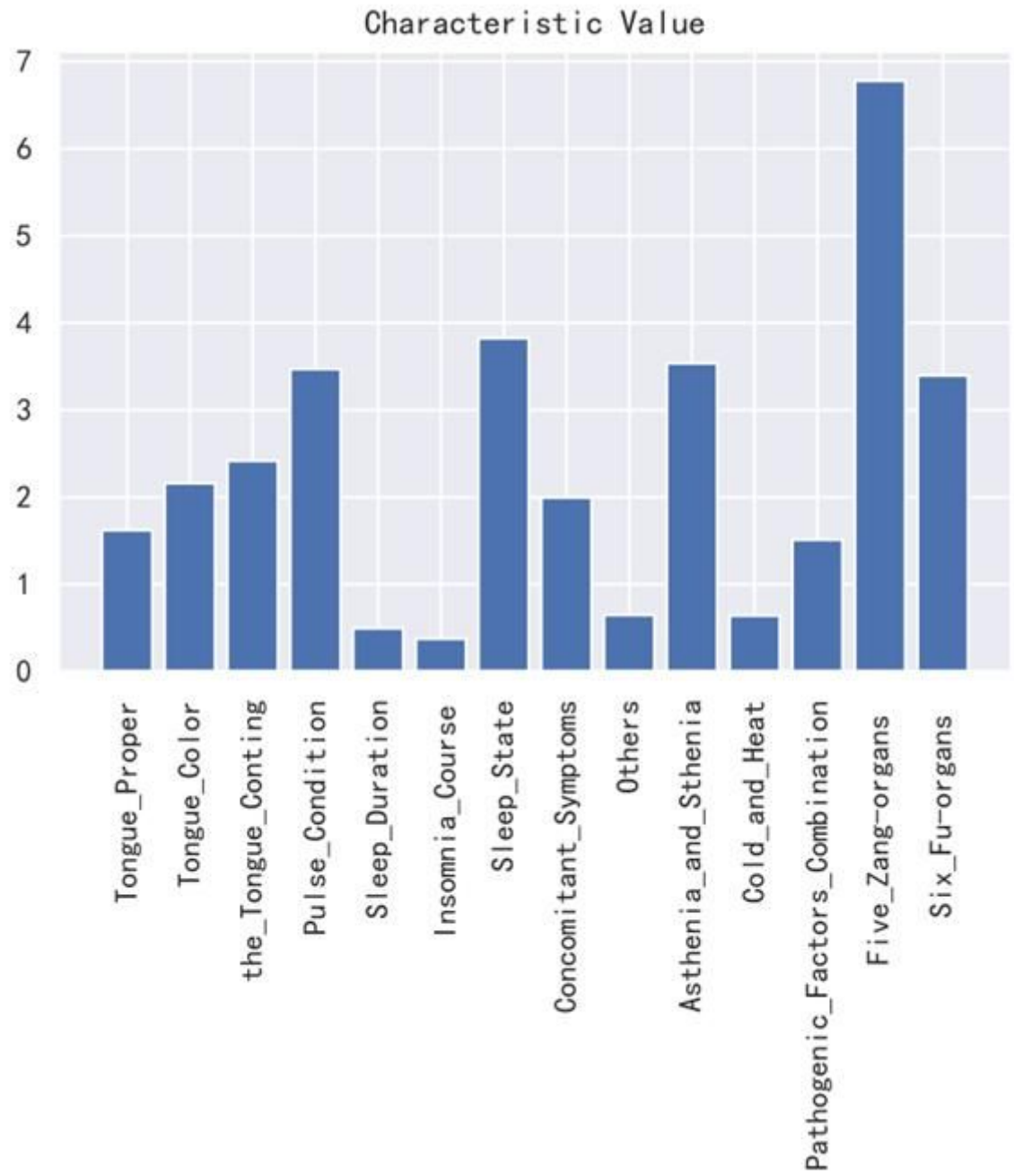

Figure 8

Transformed eigenvalues obtained by using random forest model.