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 2nd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-39359/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

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

Yuqi Tang¹, Dongdong Yang^{1*}, Zechen Li², Yu Fang¹, Shanshan Gao¹

Correspondence author: Dongdong Yang

(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 was conducted by random forest.

Results: Analysis of the four parts revealed some significant relationships. 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

1 differentiation was 0.85. Furthermore, the syndrome differentiation of five zang-organs was proven
2 to be the most significant parameter after the features extraction by the random forest.

3 **Conclusions:** The results indicate that the machine learning method can be adopted to study the
4 dominant diseases of TCM to explore the crucial rules of the diagnosis and the treatment.

5

6 **Keywords**

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

8

9 **Background**

10 The application of TCM can be traced back 25 and treatment ideas, the non-dependence of
11 to thousands of years[1]. In spite of the fact that 26 treatment drugs and the diversity of treatment
12 TCM is still regarded as the complementary and 27 schemes, etc. Unlike western medicine,
13 alternative therapy in the field of modern 28 which is based on rigorous scientific trials to
14 medicine, it can hardly be ignored that TCM has 29 support the diagnosis and treatment, most of
15 attracted widespread attention in recent years 30 TCM diagnoses are relied on the experience
16 due to its unique personalized treatment scheme 31 of doctors to get comprehensive and
17 and the outstanding treatment effect on some 32 personalized treatment strategies.
18 dominant diseases[2-3]. Insomnia is one of the 33 Consequently, TCM is considered as an
19 dominant diseases of TCM. It has been proven 34 empirical medicine as well. Nonetheless, it
20 that TCM has been successfully applied to the 35 should be noted that a set of core theories of
21 treatment of insomnia in the medical field[4-5]. 36 TCM have been established since the
22 Compared with the western medicine in the 37 beginning of the TCM development.
23 treatment of insomnia, the advantages of TCM 38 Subsequently, the core theories of TCM have
24 treatment are the personalization of diagnosis 39 been developed into the TCM prescription,

1 acupuncture, meridians and other theories[6]. 23 whole diagnosis and treatment process is not
 2 Moreover, in the long-term clinical practice, 24 only logical, but also indivisible. In the past
 3 with the constant deepening of the 25 decades, many efforts have been done to
 4 understanding of the basic theories of TCM, the 26 study this process, whereas most researches
 5 diagnosis and treatment ideas have been 27 have only focused on one part of this process.
 6 developed tremendously, and the diagnosis and 28 Zhang S et al. applied the data mining
 7 treatment standards have achieved an 29 technology to explore the drug rules of
 8 innovation as well[7]. Diagnosis and treatment 30 pulmonary fibrosis based on TCM medical
 9 ideas and treatment strategies are the critical 31 records[8]. Yu XW et al. analyzed the dose
 10 points of the clinical practice. Meanwhile, the 32 data of TCM prescriptions by optimizing the
 11 medical record data are the embodiment of 33 traditional Cheng-Church double clustering
 12 diagnosis and treatment ideas, thus worth 34 algorithm (CC)[9]. Liu y et al. adopted the
 13 exploring. The medical record of TCM is 35 data mining method to verify the TCM
 14 composed of four parts, including the basic 36 syndrome patterns of PSCI[10]. These
 15 information, the four diagnoses of TCM, the 37 researches have shown some opinions on the
 16 treatment based on syndrome differentiation and 38 diagnosis and treatment process to some
 17 the main prescription. The diagnosis and 39 extent. However, their research methods have
 18 treatment of TCM is a whole from the 40 violated the core principle of integration and
 19 information collection (including basic 41 personalization of diagnosis and treatment of
 20 information and four diagnoses) to the treatment 42 TCM, resulting that their conclusions can
 21 based on syndrome differentiation, and then to 43 hardly be applied in clinical practice[11].
 22 the development of the main prescription. The 44 Therefore, for the sake of reliability and

comprehensiveness of the research method adopted in the present paper, the analysis is carried out logically according to the sequence of TCM diagnosis and treatment, and the whole will be discussed at last. TCM diagnosis and treatment are analyzed separately at first. And then, the whole four parts is discussed. Considering that the data used in each analysis step have unique characteristics, different analysis schemes are

In recent years, the rapid development of data analysis and artificial intelligence provides 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 this paper, the above-mentioned four parts in the process of

26

27 **1 Data and method**

28 **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; ② Having difficulty in sleep maintenance, mainly manifested by easy and early to wake up;③

1 The quality of sleep is decreased, and the 23 Meanwhile, the workload is equally assigned
2 patient can hardly get into deep sleep and have 24 to the three doctors, and the cross-validation
3 multiple dreams; ④Insufficient sleep time, less 25 is implemented after all work has been
4 than 6.5 hours; ⑤With daytime symptoms, 26 completed, so as to eliminate the impact of
5 including fatigue, emotional problems, memory 27 subjectivity and artificial errors on the final
6 and attention decline, daytime sleepiness and 28 data. And then, the sample database is
7 work initiative decline, etc. The exclusion 29 established. Simultaneously, according to the
8 criteria are set as: ①The missing of the medical 30 TCM diagnosis and treatment ideas, the
9 record data is so severe that it is unable to meet 31 contents of the sample data are divide into
10 research requirement; ② The patients have 32 four parts: the basic information, the four
11 other serious organic diseases that may cause 33 diagnostic treatment, the treatment based on
12 insomnia. 34 syndrome differentiation and the main
13 In our preliminary work, 1577 outpatient 35 prescription. Each part contains several data,
14 data (from 2016 to 2020) are collected and 36 and the specific data processing steps will be
15 screened according to the above inclusion and 37 described later. In the light of the
16 exclusion criteria. Finally, only 654 outpatient 38 characteristics of the data, the methods of
17 data are selected as the research samples. Since 39 frequency analysis, association rules and
18 the selection and analysis of medical record data 40 hierarchical clustering analysis are adopted to
19 of TCM have a high demand for expertise, three 41 process and mine the data. Finally, the data
20 professional doctors of TCM are selected to 42 of the TCM diagnosis and treatment ideas
21 analyze, code and classify the medical record 43 from the four diagnosis, the treatment based
22 data information of research samples manually. 44 on syndrome differentiation and the main

1 prescription are integrally discussed by 3 specific process is illustrated in Figure 1.

2 employing the random forest algorithm. The

Figure 1 Flowchart of the research strategy designed in this paper

1 In the process of coding and classification, 3 of insomnia in China(2017)[13] and the

2 the Guidelines for the diagnosis and treatment 4 International Classification of Sleep

Disorders(ICSD2)(2005)[14] are regarded as differentiation, a complete code comparative table is shown in Table 1. Three doctors of comprehensiveness of the data. In the meantime, TCM are required to complete their work in based on the combination of TCM insomnia strict accordance with the coding table related symptoms and the syndrome

Table 1 Code comparative table of various symptoms

Project	Content	Numbe	Project	Content	Number	Project	Content	Number
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	Sleep duration	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	3		6-7 hours	8
	Others	4		Complex				
Five zang-organs	Normal	0	Pathogenic factors	Others	4	Emotional status	>7 hours	9
	Heart	1		Normal	0		Taking medicine to sleep	A
	Liver	2		Phlegm	1		Normal	0
	Spleen	3		Fire	2		Anxious	1
	Lung	4		Blood stasis	3		Fear	2
Tongue proper	Kidney	5	Sleeping status	Asthenia	4	Emotional status	Nervous	3
	Normal	0		Others	5		Restlessness	4
	Enlarged tongue	1		Normal	0		Timidity	5
	Thin tongue	2		Difficult to fall asleep	1		Irritability	6
	Teeth print on tongue	3		Dysphylaxia	2		Depressed	7
Six fu-organs	Cleft tongue	4	Others	Festless sleep	3	Concomitant symptoms	Flusteredness	8
	The vessels of sublingual	5		Hard to fall asleep again after waking up	4		Others	9
	The tongue with ecchymosis			Dreaminess	5		Normal	0
	Others	6		Others	6		Headache	1
	Normal	0		Normal	0		Dizziness	2
Six fu-organs	Stomach	1	Others	Sallow complexion	1	Concomitant symptoms	Lethargic	3
	Gallbladder	2		Flushed cheeks	2		Aversion to cold	4
	Large intestine	3		Redness of the eyes	3		Aversion to heat	5

Pulse conditions	Small intestine	4	The tongue coating	Dark lip	4	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
	Normal	0		Normal	0	Nocturia	A
	Thin	1		Yellow	1	Fatigue	B
	Wiry	2		Thin	2	Dry mouth	C
	Slippery	3		Slimy	3	Bitter taste in the mouth	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		

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.

Table 2 Comparative table of the null values and the substitute values

Items	Substitute values
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	Mixture of asthenia and sthenia
Cold and heat	Cold and heat complex
Tongue color	Others
Pathogenic factors	Others

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 "times". 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:

$$Support(X \rightarrow Y) = P(X | Y)$$

(1)

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

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

$$Confidence = \frac{P\{x|y\}}{P\{X\}} \quad (2)$$

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.

$$Lift(X \rightarrow Y) = \frac{P(Y|X)}{P(Y)} \quad (3)$$

2

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

7 1.2.4 Cluster analysis

8 At present, the cluster analysis is
9 extensively used in the medical field [16]. In
10 general, the cluster analysis can be classified
11 into two categories, one is hierarchical
12 clustering algorithm and the other is
13 agglomerative clustering algorithm. In the
14 Euclidean space, using hierarchical clustering
15 algorithm to analyze small-scale data sets can
16 achieve better results. Its basic principle is to
17 establish a hierarchical clustering tree by
18 calculating the similarity among different
19 categories of data points and adopting the
20 bottom-up aggregation strategy. Each sample
21 set in the data sets is regarded as a cluster, and
22 then the clusters with close distance are merged

23 step by step to achieve the expected number
24 of clusters.

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

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

29 (4)

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

33

34 1.2.5 Random forest

35 The random forest algorithm derived
36 from ensemble learning method is composed
37 of multiple decision trees. The random forest
38 is an extension of the classification tree and
39 the regression tree. These trees can be used to
40 model the response variables through
41 recursive partition and predict the final
42 results jointly [17]. The random forest
43 algorithm is commonly employed in data
44 classification and regression[18]. At present,
45 there are three mainstream decision tree
46 algorithms, including ID3, C4.5 and CART.

1 In the present paper, the most widely used
2 algorithm, CART, is selected to build random
3 forest algorithm model. The main function of
4 this algorithm is described below.

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

$$8 \quad Gini(D) = \sum_k \frac{|C_k|}{D} (1 - \frac{|C_k|}{D}) = 1 - \sum_k (\frac{|C_k|}{D})^2 \quad (5)$$

10
11 where C_k represents the sample subset of class k.
12 The $|C_k|$ and $|D|$ represent the size of C_k and D
13 respectively.

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

$$19 \quad D_1 = \{D | A = a\}; D_2 = \{D | A \neq a\} \quad (6)$$

22 Consequently, the $Gini(D, A)$ of set D
23 under the condition of known feature A can be

24 obtained by combining the above functions.

25 The Gini index is theoretically similar to
26 entropy, as described below.

$$27 \quad Gini(D, A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

29 (7)

31 Similar to the principle of entropy, the
32 greater the value of $Gini(D, A)$ is, the greater
33 the sample uncertainty is. Taking this into
34 account, the value of $Gini(D, A)$ should be as
35 small as possible when selecting the feature
36 A.

37 2 Results and discussion

38 2.1 Basic information

39 The basic information mainly consists
40 of the ID, name, clinic time, age and gender
41 of patients. Since the clinic time is not taken
42 as a factor in the screening criteria during the
43 data screening stage, the statistical results
44 may deviate from the actual situation. The ID
45 and name of patients privacy involved,
46 however, have no impact on the diagnosis

1 and treatment process . As a consequence, the
2 focus of this section is age and gender of
3 patients. Considering that the categories of age
4 and gender data are relatively few, we choose
5 frequency analysis for the data processing.
6 According to the box-plot of age distribution of
7 patients (shown in Figure 2), it can be seen that
8 the average age of patients is 47, the maximum
9 age is 79, the minimum age is 14 and the mean
10 square deviation of age is 11.
11

12
13 **Figure 2** Box-plot of the age distribution of
14 Patients.

15 Figure 3 depicts the distribution of
16 gender of patients. As illustrated in this
17 figure, the proportion of female patients is
18 significantly higher than that of male
19 patients.

20
21
22 **Figure 3** Pie chart of the gender distribution of
23 patients.

24 25 **2.2 Four diagnostic methods**

Four diagnostic methods include inspection (observation), auscultation and olfaction (listening and smelling), interrogation (inquiring 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” refers particularly to the method that doctors use index finger, middle finger and ring finger 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).

Figure 4 Classification of four diagnostic methods

1 Based on the smallest unit of classification, 10 TCM, and considering that there are too
2 the method of association rules is applied to 11 many null values in some of the smallest
3 study. Considering that the basic information is 12 classification units, we attempt to use two
4 also a part of TCM interrogation and may have 13 methods to analyze the association rules for
5 an effect on the diagnosis and treatment process 14 the combination of the smallest units (the
6 of the diseases, and therefore the basic 15 combination items are in brackets below), so
7 information is included in the four diagnostic 16 as to minimize the impact of the null values
8 parts for discussion as well. Combined with the 17 on the research results. The results are listed
9 characteristics of the four diagnostic methods of 18 in Table 3 and Table 4.

Table 3 Summary of the results using Method 1

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 course >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

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.

Table 4 Summary of the results using Method 2

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

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.

1 It can be concluded from the above tables
2 that most of the results are dominantly related to
3 gender and age, while there is no significant
4 relationship among the four diagnoses.
5 According to the analysis of the actual clinical
6 experience, the above results have no
7 remarkable guiding significance for clinical
8 practice. Nevertheless, two innovative research
9 directions can be found based on these results.
10 On the one hand, this research can be explored
11 deeply through expanding the sample size and
12 using other methods to find the internal
13 relationship of the four diagnoses. On the other
14 hand, there is no obvious external connection
15 among the data, but these data have statistical
16 significance. These data can be used for the
17 epidemiological study of TCM on condition that
18 the sample size is large enough.

19 20 **2.3 Treatment based on syndrome differ** 21 **entiation.**

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

30 The treatment based on syndrome
31 differentiation is a comprehensive analysis
32 by doctors in the process of diagnosis and
33 treatment of TCM, and its judgment criteria
34 are derived from the objective medical record
35 information including the four diagnoses. In
36 essence, the treatment based on syndrome

1 differentiation is a high generalization of the 23 doctors of TCM, and then classified and
 2 causes of different patients according to the 24 coded according to the above four significant
 3 principle of personalized treatment. In the light 25 syndrome differentiation points. It is worth
 4 of the logic of TCM syndrome differentiation, 26 mentioning that the organs syndrome
 5 the treatment based on syndrome differentiation 27 differentiation includes heart, liver, spleen,
 6 can be divided into three parts, including the 28 lung, kidney, gall bladder, stomach, small
 7 eight principal syndrome differentiation, the 29 intestine, large intestine, bladder and the
 8 organs syndrome differentiation and the 30 triple burner; the syndrome differentiation of
 9 meridian syndrome differentiation. Further, 31 asthenia and sthenia consists of asthenia
 10 these three parts can be separated into several 32 syndrome and sthenia syndrome; the
 11 items. The previous studies have either skipped 33 syndrome differentiation of cold and heat is
 12 this process directly or determined the 34 composed of cold syndrome and heat
 13 syndrome differentiation category only based on 35 syndrome; the pathogenic factors include
 14 the experience description of doctors, which 36 phlegm, fire, blood stasis and asthenia. The
 15 were too empirical. Based on the characteristics 37 above-mentioned 19 syndrome
 16 of insomnia in TCM, this paper focuses on four 38 differentiation factors constitute the section
 17 significant syndrome differentiation points, 39 of treatment based on syndrome
 18 namely the syndrome differentiation of asthenia 40 differentiation of the insomnia sample data
 19 and sthenia, the syndrome differentiation of 41 research in this paper. To ensure the
 20 cold and heat, the syndrome differentiation of 42 objectivity of each syndrome differentiation
 21 organs and pathogenic factors. The medical 43 factor, the three TCM doctors are supposed
 22 record data is extracted by three professional 44 to collect at least two or more kinds of

1 medical record information in the classification 13 differentiation is heat syndrome; the medical
2 and coding stage of medical record data for 14 information "slippery pulse" combined with
3 determining one syndrome differentiation factor. 15 "yellow tongue" and "greasy tongue coating"
4 For instance, the medical information "wiry 16 mean that the pathogenic factors is phlegm.
5 pulse" and "irritability" can infer that the 17 Despite that each syndrome
6 syndrome differentiation factor of organs is 18 differentiation factor in each medical record
7 liver; the medical information "thin pulse" and 19 is relatively independent, there is a strong
8 "tiredness" imply that the factor of asthenia and 20 correlation among the factors. Therefore, it is
9 sthenia syndrome differentiation is asthenia; the 21 reasonable to select association rules for
10 medical information "red tongue" combined 22 analysis. Through a process of trial and error,
11 with "tidal fever" and "rapid pulse " indicate 23 the confidence is finally adjusted to 0.7, and
12 that the factor of cold and heat syndrome 24 the results are shown in Table 5.

Table 5 Summary of the results adopting association rules to analyze the syndrome differentiation factors

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: mixture of asthenia and sthenia	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: mixture of asthenia and sthenia→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: mixture of asthenia and sthenia→Cold and heat: heat	0.72	0.28	1.12

1 As can be seen from the above table, 4 between fire and excess syndrome, fire and
 2 besides the connections that can be obtained 5 heat syndrome, there are more new
 3 from the basic theories, such as the connections 6 connections. For example, the complex

1 syndrome of heart, liver, spleen, asthenia and 23 the present paper, the TCM prescription is
2 sthenia → the heat syndrome, the fire stasis 24 the research focus of this section. TCM
3 syndrome → the heart, liver. The following 25 prescription is the embodiment of clinical
4 conclusions can be drawn by analyzing the 26 practice of TCM. Choosing the appropriate
5 treatment based on syndrome differentiation 27 combinations of Chinese medicine under the
6 with association rules. On the one hand, the 28 guidance of the treatment based on syndrome
7 results can reveal the syndrome differentiation 29 differentiation not only reflects the typical
8 thoughts of TCM doctors. On the other hand, 30 thoughts of TCM, but also conforms to the
9 after applying the above methods to classify the 31 treatment method of drug combination
10 contents of treatment based on syndrome 32 therapy[21]. Having completed the process
11 differentiation, the results can reflect the 33 from the four diagnoses to the treatment
12 priority direction of syndrome differentiation of 34 based on syndrome differentiation, the
13 insomnia to a certain extent, thus having 35 doctors should determine the main
14 guiding significance for clinical practice. In the 36 prescription. And then, on the basis of the
15 further study, more research methods can be 37 main prescription, the doctors should adjust
16 adopted to verify the dominant diseases of TCM 38 the prescription properly according to the
17 and explore new syndrome differentiation rules. 39 actual situation of patients. Finally, the

18 40 treatment prescription can be obtained. Thus,
19 **2.4 Main prescription** 41 the determination of the main prescription is

20 Basically, the treatment strategy is 42 particularly significant. The main
21 composed of acupuncture, moxibustion, 43 prescription can not only prove the
22 scraping therapy and TCM prescription, etc. In 44 personalized treatment advantages of TCM,

1 but also reflect the most core treatment method 9 component of the main prescription[22-23].

2 in the clinical practice of TCM doctors. The 10 Taking the above deficiencies into account,

3 previous studies have achieved some success; 11 the less use herbs are removed from the

4 however, there are two deficiencies in their 12 statistics of the herb use frequency, thus

5 researches. First, the previous researches have 13 reducing the impact on the research of the

6 mainly focused on the frequency of herb use 14 main prescription in this paper. Table 6

7 and interrelation of the herbs. Second, there are 15 presents the correspondence between the

8 few previous researches concerned about the 16 processed data codes and herbs.

Table 6 Correspondence between the herbs and the codes

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 aucklandiae	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		

1 For the sake of reducing calculation 8 satisfactory result. According to the
 2 amount and the increasing the code execution 9 characteristics that the main prescription is
 3 efficiency, all the herbs are replaced with codes, 10 composed of a wide variety of herbs, the
 4 and then the codes are entered into the database. 11 hierarchical clustering algorithm is applied to
 5 The hierarchical clustering algorithm is 12 explore the potential classification rules in
 6 employed to analyze the small sample data set 13 the data samples of TCM. The results of the
 7 in the European space, thus obtaining a 14 analysis of the main prescription using

1 hierarchical clustering analysis are shown in 2 Figure 5.

3

4

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

8

9 The main prescriptions of the 11 Table 7, and the repeated herb combinations
10 corresponding serial number are presented in 12 in all main prescriptions are shown in Table
13 8.

Table 7 Correspondence between the main prescriptions and the serial numbers

Serial number	Main prescription
1	Spine date seed、Glycyrrhiza、Anemarrhena、Poria cocos、Ligusticum wallichii、Caulis polygوني 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 combination and the serial numbers

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

1 The above conclusions indicate that the 3 hierarchical clustering algorithm to analyze
2 desired results can be achieved by adopting the 4 the main prescriptions. The rapid acquisition

1 of the main prescription of TCM is beneficial 17 accordance with the research process
2 for the study of the combination rules of TCM, 18 designed in the previous section(in Figure 1),
3 but also lays a solid foundation for the overall 19 the random forest algorithm is adopted to
4 study of the diagnosis and treatment of the 20 establish the model. Simultaneously, the data
5 dominant diseases of TCM. In order to facilitate 21 sets collected from four diagnoses, treatment
6 the further discussion on the overall diagnosis 22 based on syndrome differentiation, and the
7 and treatment idea, we code the main 23 main prescription of TCM are put into the
8 prescriptions of TCM and enter the codes into 24 model for cross-validation. Consequently, the
9 the database. 25 corresponding accuracy can be obtained. In

10 **2.5 Diagnosis and treatment idea**

11 In the discussion of the aforementioned 26 the meantime, for the purpose that the
12 four parts, the four parts of TCM diagnosis and 27 internal relationship of TCM diagnosis and
13 treatment ideas are studied successively, so as 28 treatment ideas can be explored deeper, this
14 to reveal the internal relationship and related 29 section is divided into two processes for
15 research methods of each part. This section 30 further discussion. These two process are
16 discusses the four parts as a whole. In 31 illustrated in Figure 6.

Figure 6 Flowchart of the diagnosis and treatment idea of TCM

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.

1 It is worth noting that five zang-organs, six 9 data processing, the five zang-organs
2 fu-organs and pathogenic factors each contains 10 combination, six fu-organs combination,
3 several syndrome differentiation factors, which 11 pathogenic factors combination and main
4 are randomly combined in the medical record 12 prescription combination are coded and put
5 sample data. In addition, in the actual outpatient 13 into the database. For the sake of
6 service, the prescription made by the doctor for 14 presenting the accuracy more intuitively, the
7 patients commonly includes at least one main 15 method of confusion matrix is carried out in
8 prescription. Therefore, in order to facilitate

1 this paper. The confusion matrix results are 2 shown in in Figure 7.

Figure 7 Confusion matrix

The process 1 is shown in Figure 7(A-E). The process 2 is presented in Figure 7(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.

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.

Table 9 Accuracy of applying the random forest algorithm model

Items	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

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.

Figure 8 Transformed eigenvalues obtained by using random forest model

As can be seen from the Figure 8, the most significant parameter affecting the

judgment results is the syndrome differentiation of five zang-organs , followed by the sleep status, the pulse condition, 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 condition 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 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 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 comparison, so as to further investigate the feasibility of machine learning

1 in the research of TCM diagnosis and treatment.

2

3 **3 Conclusions**

4 In this paper, the machine learning
5 methods are adopted to deeply mine and
6 analyze the medical record data of the dominant
7 diseases of TCM, providing a new strategy and
8 direction for the development of the clinical
9 diagnosis and treatment technology of TCM.
10 The focus of this study is to analyze the
11 diagnosis and treatment process of the TCM
12 dominant diseases which includes the
13 acquisition of the patients' condition
14 information through using four diagnostic
15 methods, and the flexible application of the
16 syndrome differentiation methods to develop
17 the treatment plan and select the main
18 prescription. First of all, based on the
19 standardized data sets, the frequency analysis,
20 the association rules and the hierarchical cluster
21 analysis are applied to analyze the basic
22 information, the four diagnostic methods, the

23 syndrome differentiation and treatment and

24 the main prescription of TCM in the

25 diagnosis and treatment process. And then,

26 the random forest algorithm model is

27 established for the discussion of the whole

28 diagnosis and treatment process. Finally, the

29 extraction of the eigenvalues is conducted by

30 using the random forest algorithm model.

31 The results demonstrate that the most

32 significant parameter affecting the diagnosis

33 and treatment process is the syndrome

34 differentiation of five zang-organs, followed

35 by sleep status, pulse condition, the

36 syndrome differentiation of asthenia and

37 sthenia and the six fu-organs syndrome

38 differentiation.

39 The results indicate that the research

40 strategy established in this paper can

41 efficiently filter the unessential diagnosis and

42 treatment information, thus helping TCM

43 doctors to quickly and efficiently obtain

44 valuable information and crucial rules from a

1 substantial number of medical record data. 5 the research strategy established in this paper
 2 Furthermore, since the research process, the 6 can be applied to further investigate the
 3 data collection and the data analysis methods 7 diagnosis and treatment rules of other TCM
 4 designed in this paper are highly standardized, 8 dominant disease

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. doi:10.1007/s11684-011-0126-x
- 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 Post stroke 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 Disorders 697 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. doi:10.1109/RBME.2010.2083647
 - 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 Environ 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. Published 2015 Mar 23.
 - 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. Published 2018 Jul 4.

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.

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 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 idea of TCM 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 The process 1 is shown in Figure 7(A-E). The 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