Original Paper

Symptom Distribution Regularity of Insomnia: Network and Spectral Clustering Analysis

Fang Hu[1], PhD; Liuhuan Li[1], BSc; Xiaoyu Huang[2], BSc; Xingyu Yan[1], BSc; Panpan Huang[2], PhD

[1]College of Information Engineering, Hubei University of Chinese Medicine, Wuhan, China

[2]College of Basic Medicine, Hubei University of Chinese Medicine, Wuhan, China

Corresponding Author:
Panpan Huang, PhD
College of Basic Medicine
Hubei University of Chinese Medicine
No. 16 Huangjiahu West Road
Hongshan District
Wuhan, 430065
China
Phone: 86 15327193915
Email: panpanhuang@hbtcm.edu.cn

Abstract

Background: Recent research in machine-learning techniques has led to significant progress in various research fields. In particular, knowledge discovery using this method has become a hot topic in traditional Chinese medicine. As the key clinical manifestations of patients, symptoms play a significant role in clinical diagnosis and treatment, which evidently have their underlying traditional Chinese medicine mechanisms.

Objective: We aimed to explore the core symptoms and potential regularity of symptoms for diagnosing insomnia to reveal the key symptoms, hidden relationships underlying the symptoms, and their corresponding syndromes.

Methods: An insomnia dataset with 807 samples was extracted from real-world electronic medical records. After cleaning and selecting the theme data referring to the syndromes and symptoms, the symptom network analysis model was constructed using complex network theory. We used four evaluation metrics of node centrality to discover the core symptom nodes from multiple aspects. To explore the hidden relationships among symptoms, we trained each symptom node in the network to obtain the symptom embedding representation using the Skip-Gram model and node embedding theory. After acquiring the symptom vocabulary in a digital vector format, we calculated the similarities between any two symptom embeddings, and clustered these symptom embeddings into five communities using the spectral clustering algorithm.

Results: The top five core symptoms of insomnia diagnosis, including difficulty falling asleep, easy to wake up at night, dysphoria and irascibility, forgetful, and spiritlessness and weakness, were identified using evaluation metrics of node centrality. The symptom embeddings with hidden relationships were constructed, which can be considered as the basic dataset for future insomnia research. The symptom network was divided into five communities, and these symptoms were accurately categorized into their corresponding syndromes.

Conclusions: These results highlight that network and clustering analyses can objectively and effectively find the key symptoms and relationships among symptoms. Identification of the symptom distribution and symptom clusters of insomnia further provide valuable guidance for clinical diagnosis and treatment.

(JMIR Med Inform 2020;8(4):e16749) doi: 10.2196/16749

KEYWORDS

insomnia; core symptom; symptom community; symptom embedding representation; spectral clustering algorithm

Introduction

Background

Insomnia is a subjective complaint of a sleep disorder in which the patient has difficulty falling asleep or remaining asleep as long as desired. Insomniacs usually have low energy, less concentrating power, reduced appetite, and mood swings, leading to low performance throughout the day at work [1]. Approximately 16% of the population is reported to suffer from insomnia [2]. Clinical research has shown that traditional Chinese medicine (TCM) can be successfully applied in the treatment of insomnia [3,4]. However, the evaluation criteria of TCM diagnosis and treatment of insomnia remain unexplored. The most fundamental reason for this lack is that the clinical manifestations of insomnia are complicated and diverse; therefore, TCM physicians have difficulties in accurately extracting the core symptoms to carry out effective treatment according to clinical characteristic categories.

Machine learning, a subset of artificial intelligence and a data-oriented approach, has attracted substantial attention from various domains [5,6]. Researchers have already proposed a huge number of algorithms and models referring to machine learning to discover the hidden relationships between entities from different research fields [7,8]. TCM datasets have characteristics of "big data," particularly with respect to the complex relationships among diseases, syndromes, symptoms, prescriptions, herbs, diagnosis, and treatment [9]. As the key clinical manifestations of patients, symptoms play a significant role in clinical diagnosis and treatment, which evidently have their underlying TCM mechanisms. There are frequently multiple interrelated symptoms under the same subgroup. A symptom network reflects the macroscopic law of the dynamic process of complex symptoms under the influence of certain driving forces. In recent decades, several researchers have applied various machine-learning approaches to discover the potential regulations for treating insomnia. Ahuja et al [10] applied 15 machine-learning algorithms and took 14 leading factors into consideration for predicting insomnia. The results of this analysis showed that insomnia primarily depends on vision problems, mobility problems, and sleep disorder. Park et al [11] developed 3 prediction models for sleep quality using machine-learning techniques to uncover the relationships between sleep quality and sleep-related factors. The results suggested that morning activity, and exposure to total and outside light during daytime are important contributors to sleep quality. Based on the Bayesian belief network model, Seixas et al [12] assessed the sleep duration and physical activity profiles that provided the lowest diabetes prevalence among black and white subjects. Hu et al [13,14] discovered the core symptoms and symptom distribution rule of insomnia using a network analysis method. Li et al [15] explored suitable preprocessing methods for analysis of TCM clinical data based on a prospective study on patients with insomnia treated according to syndrome differentiation. Weng et al [16] determined the frequency of each herb and association rules among the herbs for insomnia using data mining methods.

With continuous development of artificial intelligence, heterogeneous information network [17] and graph embedding [18] can be conducted to construct a medical network and train the various medical node embeddings for in-depth analysis of TCM data, including analysis of the molecular mechanisms of symptoms [19], herb target prediction [20], and disease comorbidity patterns [21]. Yang et al [22] proposed a heterogeneous network embedding representation algorithm to construct a heterogeneous symptom-related network, which was applied to obtain the low-dimensional vector representation of symptom nodes. This model was used to predict disease genes with high performance and obtained better results than other well-known disease gene prediction algorithms. Wang et al [20] presented an herb target interaction network approach for novel herb target prediction mainly relying on symptom-related associations. The above studies helped to effectively discover the relationships among disease mechanisms, symptoms, herbs, targets, ingredients, genes, and related factors; however, the critical factors of syndrome differentiation and treatment, and their corresponding relationships require further study. In particular, the most effective methods for exploring the key factors and relationships in TCM data, and to support the clinical diagnosis and treatment remain unclear.

Objectives

In this study, we explored the potential regularity of symptoms for diagnosing insomnia using complex network and machine-learning approaches. After constructing the symptom network with specific criteria, we identified the most important symptom nodes using four node importance evaluation metrics. Using the node-embedding technique [23,24], we acquired each symptom node embedded in the symptom network, and constructed the specific symptom vocabulary with the digital formation of vectors. Further, we divided the symptoms into several communities through similarity calculations between any two symptom embeddings using the spectral clustering algorithm. Finally, we obtained the core symptoms and symptom clusters, and then summarized the symptom distribution rule of insomnia. Compared to previous studies, we combined the complex network with a machine-learning approach to find the key symptoms and their corresponding symptom distribution rule. This study will provide a novel exploratory analysis method to discover clinically relevant information from TCM data.

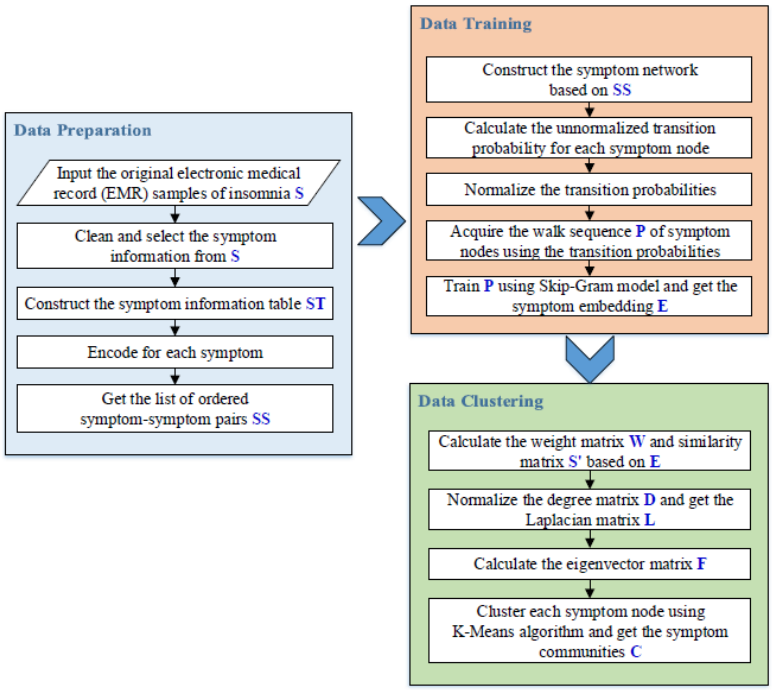
Methods

Data Extraction

The analysis dataset of insomnia was extracted from the hospital information system at Guo Yi Tang Affiliated Hospital of Hubei University of Chinese Medicine (Wuhan, Hubei, China). The inclusion criteria for record selection were patients diagnosed with typical symptoms of insomnia (sleep disorder is the main symptom and the other symptoms are secondary to insomnia), aged 14-70 years, and insomnia occurring between 1 month and 30 years. The exclusion criteria were noncollaborators, including those unable to adhere to treatment or any noncompliance that would affect data collection and efficacy evaluation, and pregnant women or terminally ill patients.

Based on these criteria, we extracted 807 effective outpatient electronic medical records (EMRs) as the research data. Through analyzing the theme data, we cleaned the raw data and selected some significant features, including syndromes and their corresponding symptoms, and then formed the analysis dataset of insomnia.

Figure 1. Flowchart of data processing.



Steps of Data Processing

A summary of the data processing for insomnia is outlined in Figure 1. We divided the data processing into three steps: data preparation, data training, and data clustering.

In the first step, we obtained the original EMRs dataset *S* from the hospital information system, cleaned and selected the symptom information from *S*, and then constructed the symptom information table *ST*. After encoding each symptom, the list of ordered symptom-symptom pairs *SS* was acquired.

In the second step, we constructed the symptom network based on *SS*, calculated the transition probability for each symptom node, and normalized the probabilities to acquire the walk sequence *P* of symptom nodes. After training *P* based on the Skip-Gram model [25], we obtained the symptom embeddings *E*.

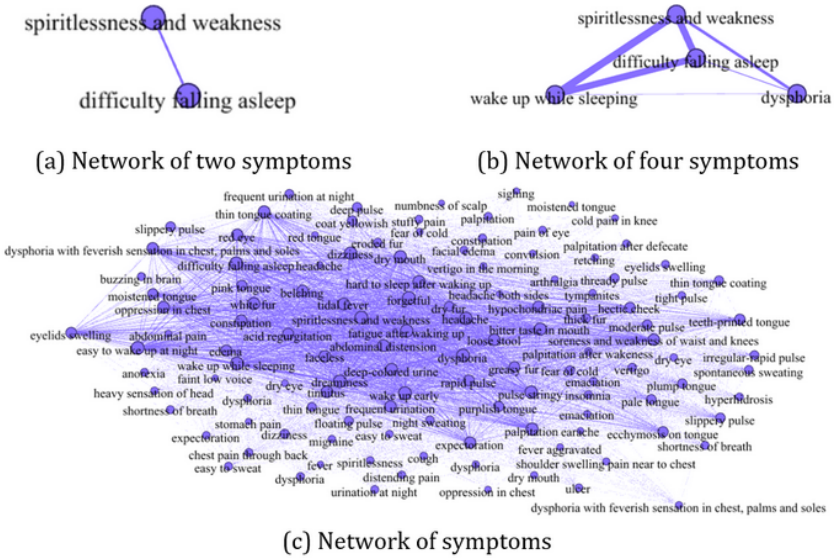
In the third step, we calculated the weight matrix *W* and similarity matrix *S′* based on the symptom embeddings *E*. From the degree matrix *D* and the Laplacian matrix *L*, we obtained the eigenvector matrix *F*. After clustering *F* using the K-means algorithm, the symptom communities *C* were acquired.

Construction of the Symptom Network Model

Based on complex network theory [26,27], we constructed the insomnia symptom network *G(V,E)*, where *V* is the node set of symptoms and *E* denotes the edge set between any two symptoms. The rules of symptom network construction were as follows: each symptom in the records was considered a node in the network, the connection between any two symptoms co-occurring in the same diagnosis was considered an edge, and the weight of an edge was considered as the co-occurrence frequency of any two symptoms.

The construction process of the insomnia symptom network based on these rules is schematically outlined in Figure 2. As shown in Figure 2a, we constructed a network with two symptom nodes, *spiritlessness and weakness* and *difficulty falling asleep*, and denoted an edge representing these two symptoms co-occurring in the same diagnosis. During development, two other symptom nodes, *wake up while sleeping* and *dysphoria*, and their corresponding weighted edges were added to the network, as shown in Figure 2b. Finally, we acquired an undirected and weighted symptom network of insomnia including 164 nodes and 10,244 edges, as shown in Figure 2c.

Figure 2. Construction process of the symptom network.



Evaluation Metrics of Node Centrality

For complex networks, several evaluation metrics of node centrality are typically used to identify the core nodes [28]. The representative metrics include degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality, which can reflect the node centrality (also called node importance) from different aspects. Degree centrality reflects the direct influence and the acquiring information ability of one node [29], closeness centrality reflects the distance properties between one node and other nodes [29], betweenness centrality measures the proportion of the shortest paths through one node [29], and eigenvector centrality represents the importance of one node comprehensively considering the importance of its neighbor nodes [30]. The equations of these four evaluation indices are as follows:

Degree centrality:

$$C_d(v) = \frac{\deg(v)}{N-1}$$

Betweenness centrality:

$$C_b(v) = \frac{\sum_{s \neq v \neq t \in V} \frac{\delta_{st}(v)}{\delta_{st}}}{(N-1)(N-2)/2}$$

Closeness centrality:

$$C_c(v) = \frac{\sum_{t, v \in V} d_G(v, t)}{N-1}$$

Eigenvector centrality:

$$C_e(v) = \lambda^{-1} \sum_{t=1}^{N} a_{vt} e_t$$

The complex network is denoted as *G(V,E)*, where *V* is the set of nodes and *E* is the set of edges. In the equation of degree centrality, *deg* (*v*) is the degree of node *v* and *N* is the number of nodes. In the betweenness centrality, δ_{st} is the number of the shortest paths from node *s* to node *t*, and δ_{st} *(v)* is the number of shortest paths through node *v*. In the closeness centrality equation, $d_G(v,t)$ is the shortest path from node *v* to node *t*. In the eigenvector centrality, *A* represents the adjacent matrix of a network; if there is an edge between node *v* and node *t*, a_{vt}=1, otherwise a_{vt}=0. λ_1, λ_2,…, λ_N are the eigenvalues of *A*, and e_t is the eigenvector of λ_t.

Pearson Correlation Coefficients of Symptoms

The Pearson correlation coefficient, sometimes called the Pearson product-moment correlation coefficient, is a measure of the linear correlation between two variables [31,32]. It has a value between –1 and +1, where +1 indicates a complete positive linear correlation, 0 is no linear correlation, and –1 is a complete negative linear correlation. The definition of Pearson correlation coefficient *r* is as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where *n* is the sample size; x_i and y_i are the individual sample points indexed with *i*; $\bar{x}$ is the sample mean represented as:

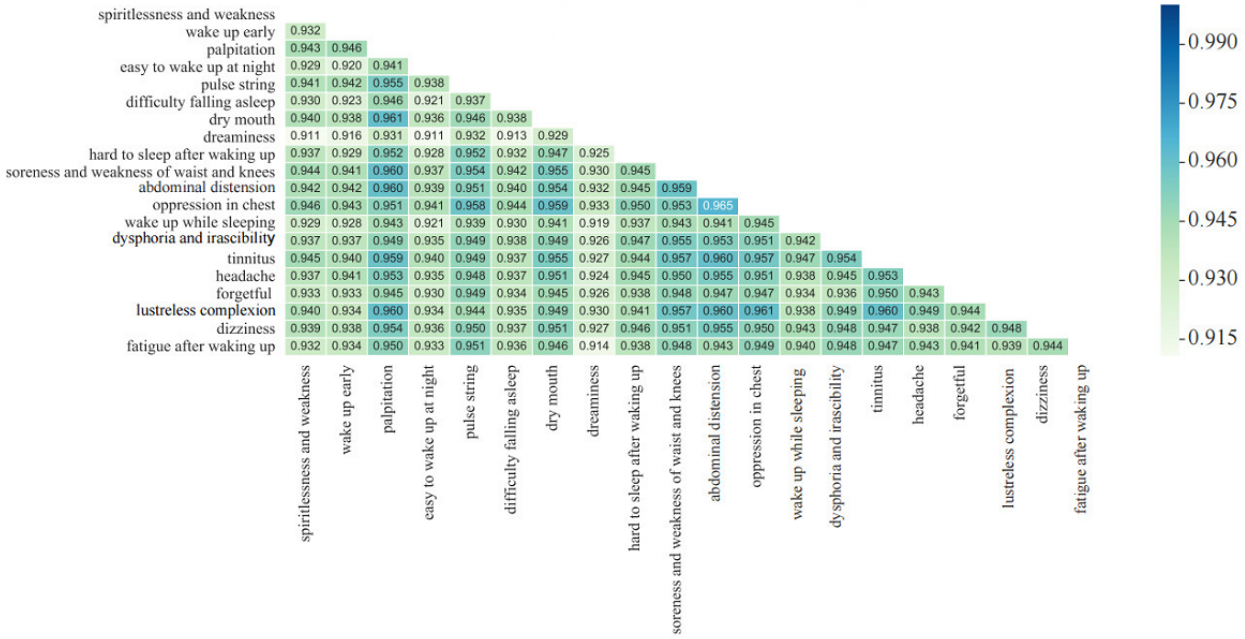
$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

and analogously for $\bar{y}$.

We calculated the Pearson correlation coefficients between any 2 of the top 20 core symptom nodes from the symptom network. The relative heatmap is provided in Figure 3, in which the

strengths of correlation values are represented using different color shading.

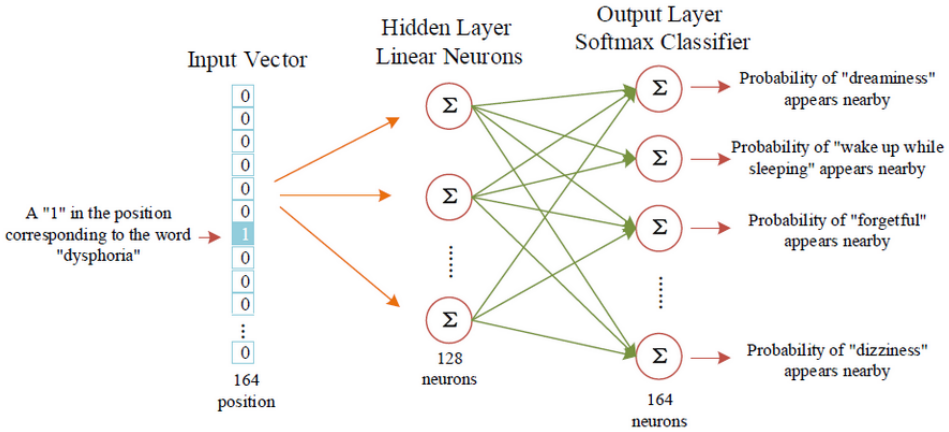
Figure 3. Pearson correlation coefficients between any two symptoms.



Training the Symptom Embeddings

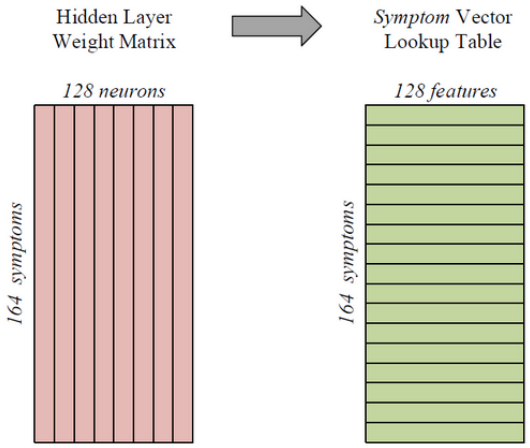
Based on the matrix of the symptom network, we use the Skip-Gram model [25] to train the insomnia symptom embeddings (also called symptom vectors). We first built a vocabulary of 164 insomnia symptom terms. We represent an input symptom term such as *dysphoria* as a one-hot vector. This vector will have 164 components (one for every symptom in our vocabulary), and we placed "1" in the position corresponding to the symptom *dysphoria* and "0" in all other positions. The output of the network is a single vector containing 128 components. For each symptom in our vocabulary, the probability of randomly selecting a nearby symptom was calculated. The neural network model for training the symptom embeddings is outlined in Figure 4. In this model, we set the input layers as the 164 one-hot symptom vectors, the number of neurons in the hidden layer as 128, and the activation function in the output layer as the softmax function. Therefore, when evaluating the trained network on an input symptom one-hot vector, the output vector will be a probability distribution (ie, a series of floating point values rather than a one-hot vector). Consequently, we can obtain the probabilities of the symptoms such as *dreaminess*, *wake up while sleeping*, *forgetful*, and *dizziness* appearing close to the symptom *dysphoria* in the network.

Figure 4. Skip-Gram model of symptoms.



After training the model as shown in Figure 4, we acquired the weight matrix (ie, the symptom embeddings with 128 features) in the hidden layer. This weight matrix has 164 rows (one for each symptom in our vocabulary) and 128 columns (one for every hidden neuron). The symptom embedding lookup table is obtained from the weight matrix in the hidden layer as shown in Figure 5.

Figure 5. Representation of symptom embeddings.



Clustering the Symptom Embeddings

To find the rule of symptom distribution and the symptom clusters of insomnia, we used the spectral clustering algorithm [33,34]—as a representative community detection algorithm used in complex networks—to divide the symptom network with 164 nodes and 10,244 edges into real communities. A community comprises one group or cluster of nodes in which the links between nodes are densely connected to each other but are sparsely connected with other communities [35].

We calculated the similarity values between any two symptom embeddings and divided the symptoms with high similarity values into the same community. The clustering process is as follows: we constructed the weight matrix W (ie, similarity matrix) through calculating the specific distance between two arbitrary symptom nodes v_i and v_j, obtained the degree matrix D, calculated the Laplacian matrix ($L=D–W$), and obtained the normalized Laplacian matrix L'. We then found the first k minimum eigenvalues and their corresponding eigenvectors of L', and constructed the eigenmatrix F using these eigenvectors. F was clustered using the K-means algorithm to finally acquire the symptom clusters of insomnia.

Results

Core Symptom Analysis

We used four evaluation metrics to calculate the different centrality values of each node in the symptom network, and display the top 20 significant symptoms of 164 nodes in Table 1. The plots for degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality are presented in Figure 6, 7, 8, and 9, respectively. The significant symptoms calculated by these four approaches were nearly identical. In particular, the degree centrality, closeness centrality, and betweenness centrality identified the same top 5 core symptoms, including *difficulty falling asleep*, *easy to wake up at night*, *dysphoria and irascibility*, *forgetful*, and *spiritlessness and weakness*. The eigenvector centrality found the same 3 symptoms *difficulty falling asleep*, *easy to wake up at night*, and *spiritlessness and weakness*, and could also find two other symptoms *wake up while sleeping* and *dreaminess*. Therefore, based on the symptom network of insomnia, the core symptoms can be identified accurately using these evaluation metrics referring to multiple aspects.

Table 1. Node centrality analysis of the symptom network[a].

No.	Symptoms	Degree	Closeness	Betweenness	Eigenvector
1	difficulty falling asleep	0.9632[b]	0.9645[b]	0.025[b]	0.2027[b]
2	forgetful	0.9325[b]	0.9368[b]	0.0204[b]	0.1997
3	dysphoria and irascibility	0.9325[b]	0.9368[b]	0.0224[b]	0.1834
4	easy to wake up at night	0.9264[b]	0.9314[b]	0.0244[b]	0.2042[b]
5	spiritlessness and weakness	0.9202[b]	0.9261[b]	0.0183[b]	0.2093[b]
6	wake up while sleeping	0.908	0.9157	0.0176	0.201[b]
7	wake up early	0.9018	0.9106	0.0176	0.1945
8	dreaminess	0.8834	0.8956	0.0129	0.225[b]
9	dizziness	0.865	0.8811	0.0162	0.1846
10	fatigue after waking up	0.865	0.8811	0.0143	0.1705
11	pulse string	0.865	0.8811	0.0177	0.1642
12	hard to sleep after waking up	0.8589	0.8763	0.0176	0.1709
13	dry mouth	0.8528	0.8717	0.0163	0.1746
14	headache	0.8466	0.867	0.0131	0.186
15	palpitation	0.8282	0.8534	0.0124	0.1556
16	abdominal distension	0.7853	0.8232	0.0115	0.1491
17	soreness and weakness of waist and knees	0.7669	0.8109	0.0099	0.169
18	tinnitus	0.7607	0.8069	0.0083	0.147
19	oppression in chest	0.7546	0.803	0.0096	0.1547
20	lusterless complexion	0.7239	0.7837	0.006	0.1477

[a]The top 20 symptoms are ranked in order of importance.

[b]The top 5 most important values in each column.

Figure 6. Degree centrality of symptoms.

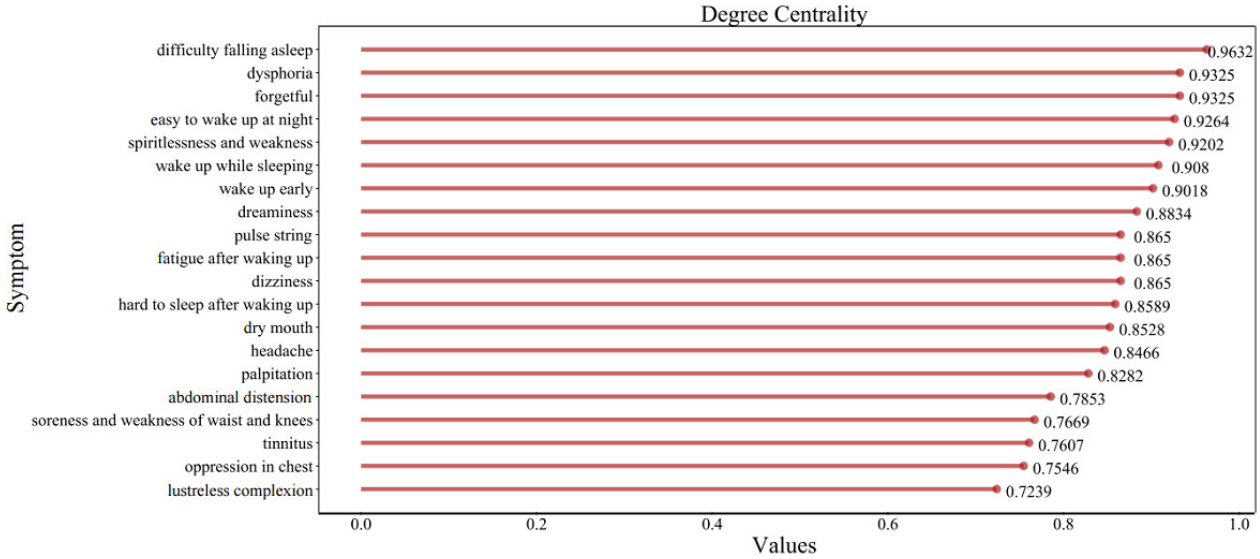


Figure 7. Closeness centrality of symptoms.

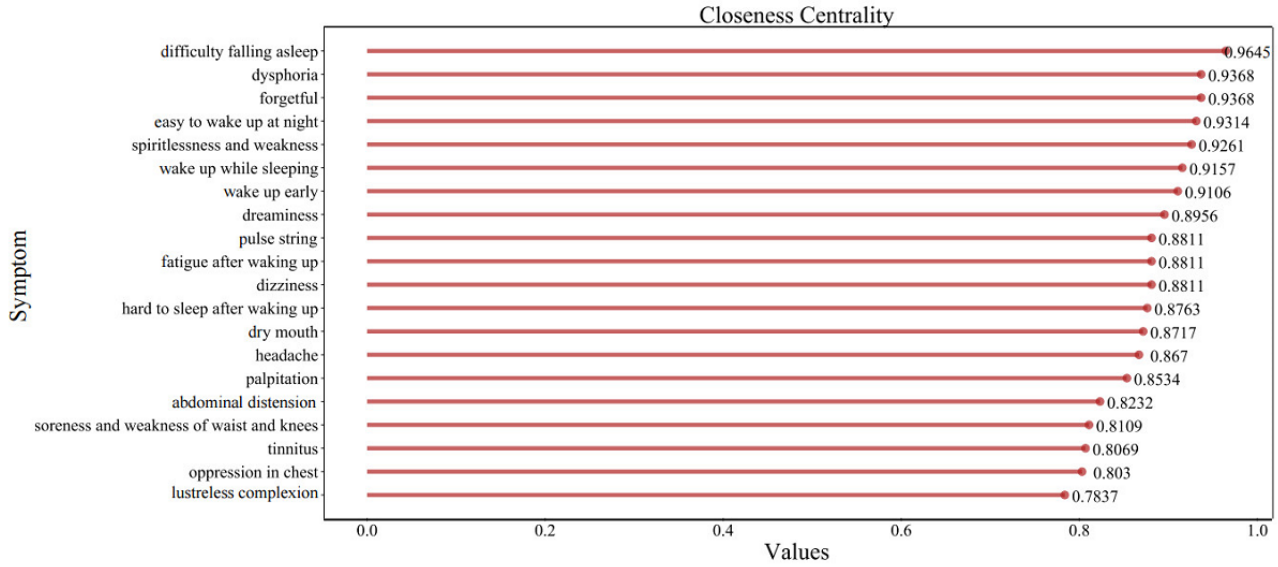


Figure 8. Betweenness centrality of symptoms.

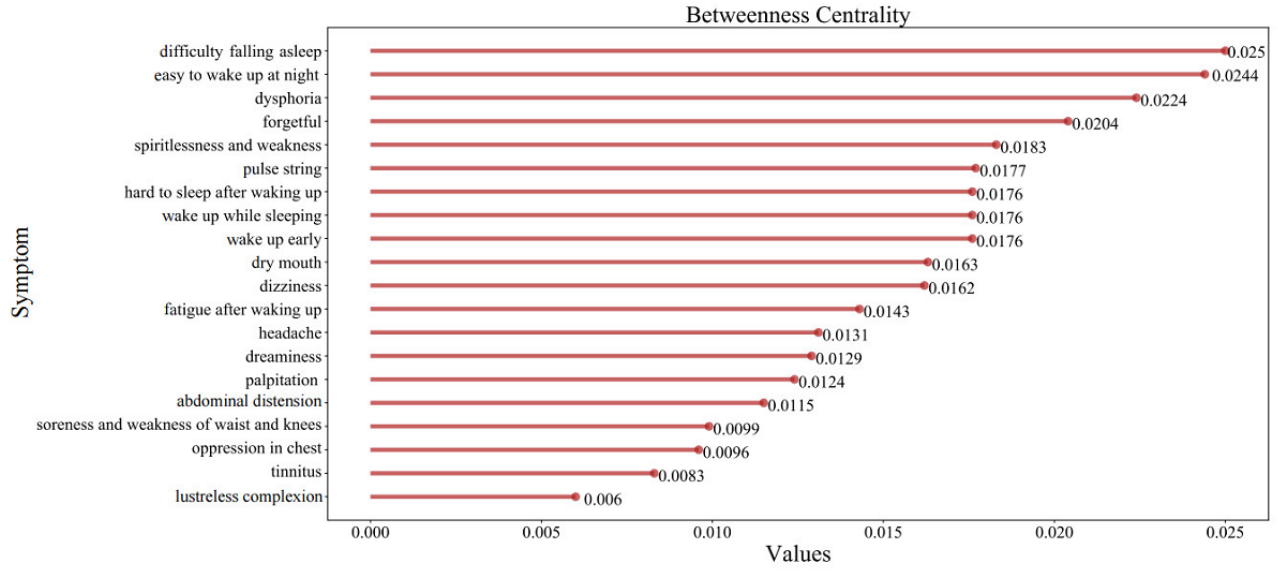
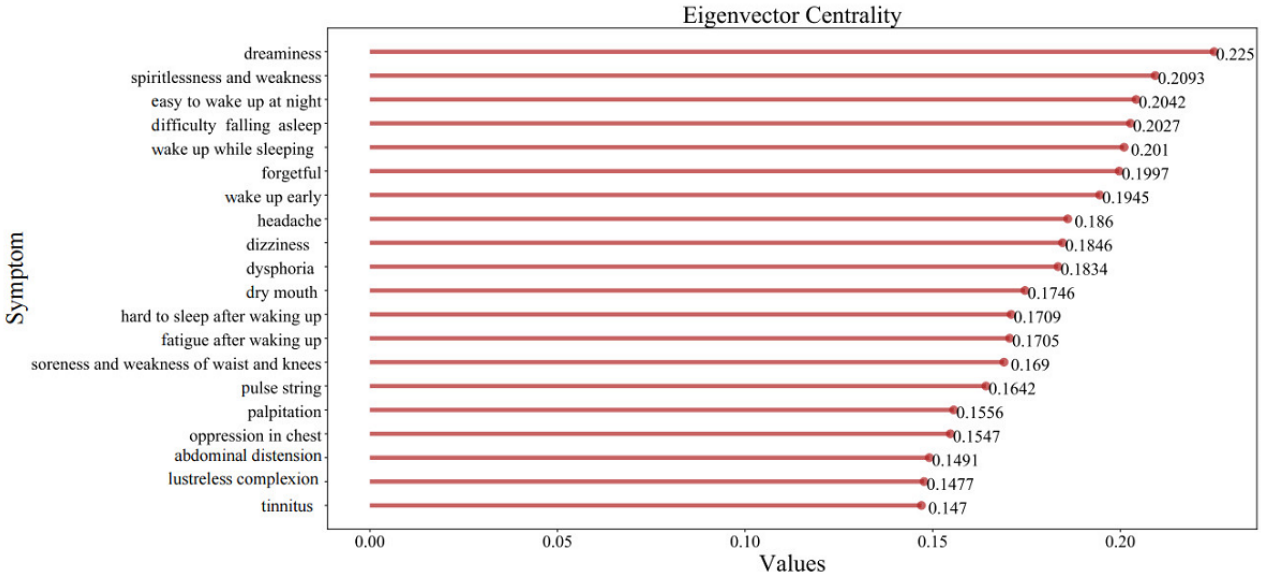


Figure 9. Eigenvector centrality of symptoms.



Symptom Correlation Analysis

Based on Figure 3, strong correlations were identified between any two of the top 20 symptoms with a range of 0.91 to 0.97. The correlation coefficient between *oppression in chest* and *abdominal distension* was 0.97, denoting that these two symptoms have the strongest correlation. A correlation coefficient of 0.96 was obtained between pairs of the following symptoms: *palpitation* and *soreness and weakness of waist and knee, pulsing string, dry mouth, abdominal distension, tinnitus, lusterless complexion; pulsing string* and *oppression in chest, tinnitus, lusterless complexion, soreness and weakness of waist and knee, abdominal distension; abdominal distension* and *tinnitus, lusterless complexion, dizziness; oppression in chest* and *tinnitus, dry mouth, lusterless complexion;* and *tinnitus* and *lusterless complexion*. These results indicate that there are strong correlations between these symptoms for the clinical diagnosis of insomnia.

Symptom Clustering Analysis

To obtain the best result of symptom distribution, we trained the symptom embeddings using the different embedding dimensions d=128 and d=164 in the node-embedding model and divided the symptom network into different communities by changing the cluster numbers (c=4 and c=5) in the spectral clustering algorithm.

The obtained symptom communities with different embedding dimensions and cluster numbers are shown in Figures 10-13. In these networks, the size of nodes denotes the degree of importance of symptoms of insomnia to the network; that is, a larger node indicates that this symptom is more important to insomnia. The size of the edges represents the co-occurrence frequencies of any two symptoms in the records. The clustering result revealed the classic symptom clusters of insomnia.

Some core symptoms such as *dry hair* in Figure 10, *frequent urination* in Figure 12, and *oppression in chest* in Figure 13 do not appear very frequently among the main complaints of patients. In addition, with regard to the disease subtypes for personalized treatment of insomnia, insomnia symptoms were only divided into four categories based on Figure 10 and Figure 12, which are too simple and cannot reflect the complexity and changeability of symptom characteristics of insomnia patients. In Figure 11, this symptom network (Figure 2) is split into five communities using the spectral clustering algorithm, which are more identical to the clinical diagnosis, as follows.

- Community 1 (green): symptoms including *spiritlessness and weakness, wake up while sleeping, fatigue after waking up, easy to wake up at night,* and *dreaminess* are divided into a community with the core symptom *difficulty falling asleep*.
- Community 2 (purple): symptoms including *dry hair, constipation, palpitation,* and *abdominal distension* are divided into a community with the core symptom *hard to sleep after waking up.*
- Community 3 (blue): the symptoms including *bitter taste in mouth, dry eye, rapid pulse, emaciation, and moderate pulse* are divided into a community with the core symptom *soreness and weakness of waist and knees*.
- Community 4 (pink): the symptoms including *purplish tongue, ulcer, earache, oppression in chest,* and *dry mouth* are divided into a community with the core symptom *pulse string.*
- Community 5 (orange): the symptoms including *expectoration, night sweating, thin tongue, floating pulse,* and *dizziness* are divided into a community with the core symptom *tinnitus.*

Figure 10. Symptom communities (d=128 and c=4).

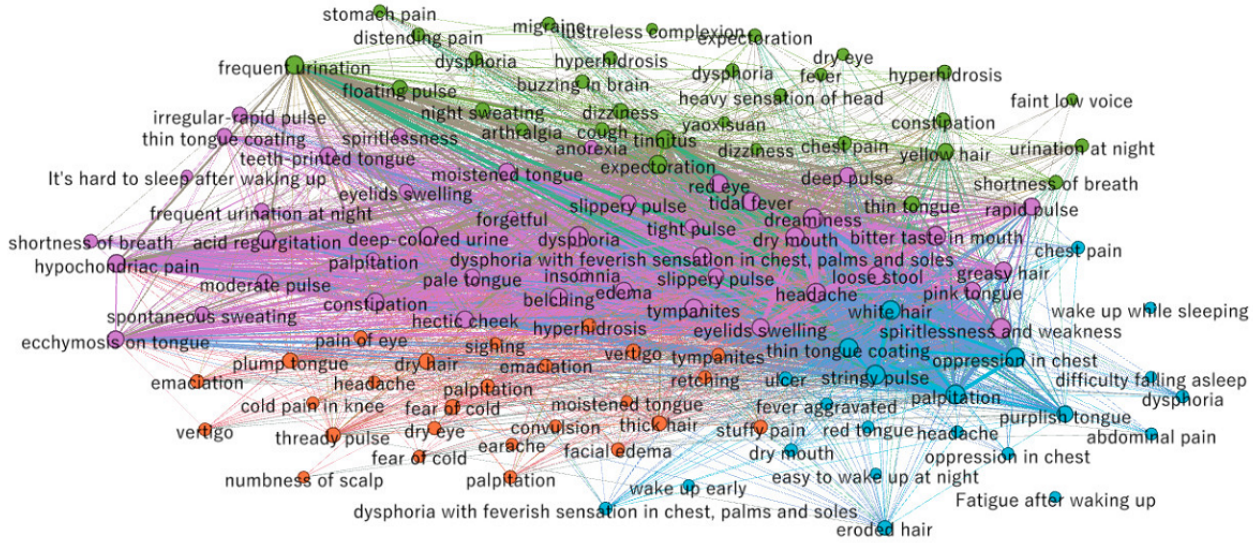


Figure 11. Symptom communities (d=128 and c=5).

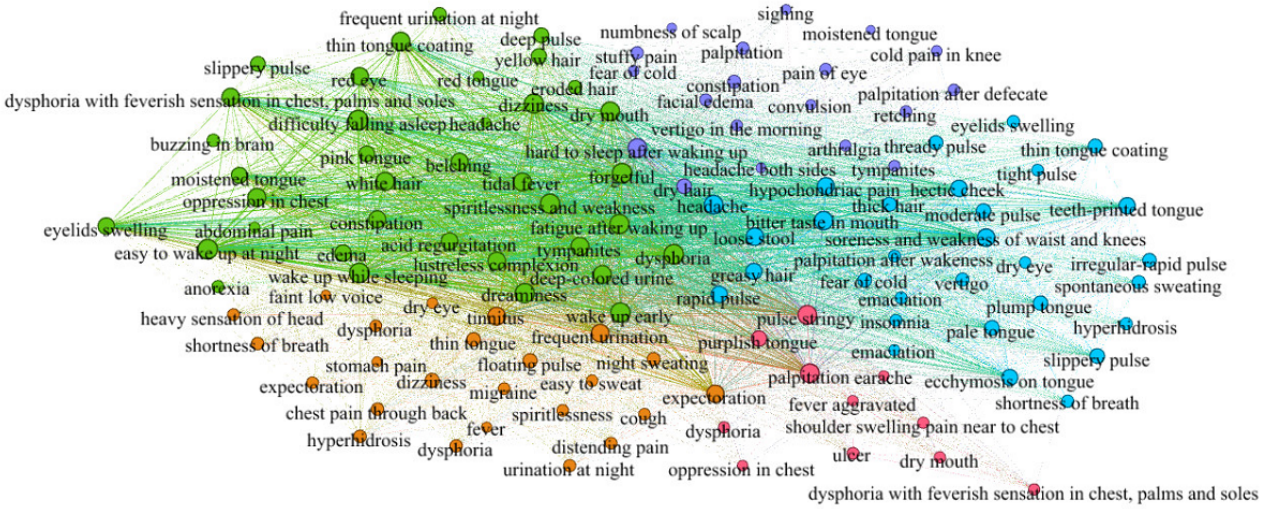


Figure 12. Symptom communities (d=164 and c=4).

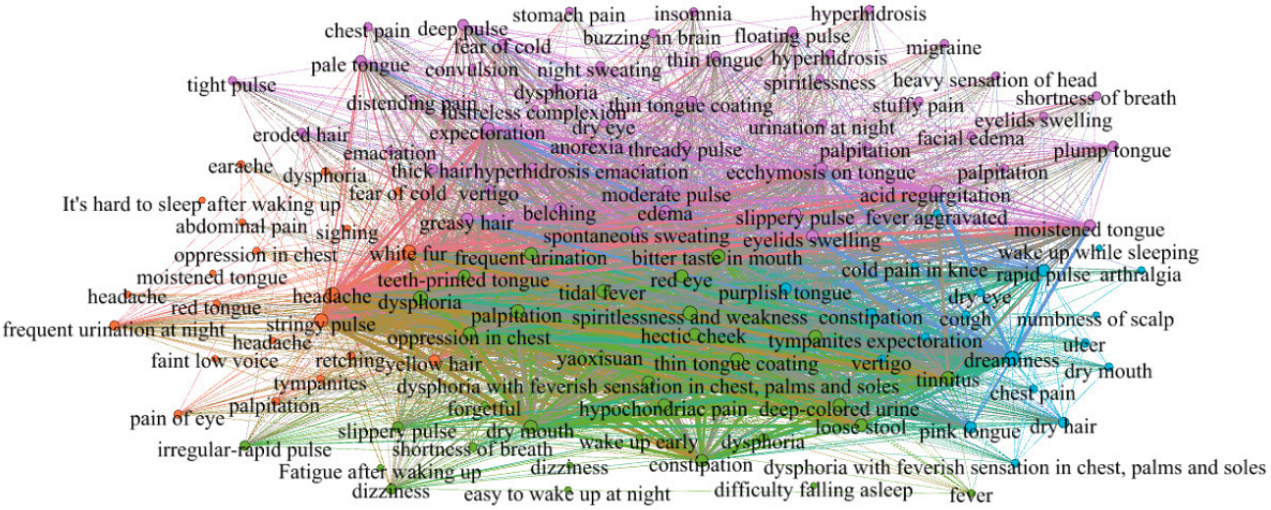
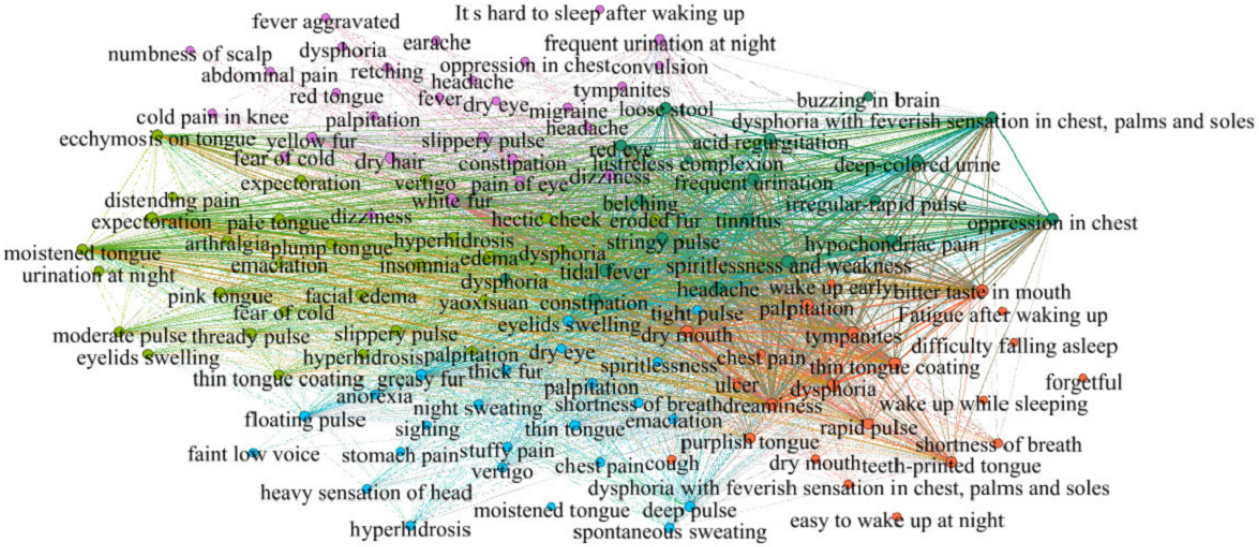


Figure 13. Symptom communities (d=164 and c=5).



Discussion

Principal Findings

In this study, we considered insomnia as a model condition, and explored the symptom distribution regularity using complex network and machine-learning approaches focusing on a node-embedding representation. We constructed the symptom network to reflect the hidden relationships between symptoms, and then identified the core symptoms using representative evaluation metrics of node centrality. Based on the symptom network, we trained the symptom vocabulary using the node-embedding technique. After clustering symptom embeddings using the spectral clustering algorithm, we acquired the insomnia symptom communities, which can reveal the symptom distribution rule. The core symptoms were identified using representative evaluation indices of node centrality such as degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality.

The results showed that the core symptoms are *difficulty falling asleep*, *easy to wake up at night*, and *dysphoria and irascibility*. Clinical research demonstrates that these symptoms always appear in the diagnosis of insomnia, and the majority of patients with insomnia have these three symptoms. According to the diagnostic criteria of International Classification of Sleep Disorders-3 in the European guidelines for the diagnosis and treatment of insomnia [36], the diagnostic criteria of chronic insomnia are: *difficulty falling asleep, difficulty maintaining sleep, getting up early, unwilling to go to bed on time, and difficulty falling asleep without intervention from parents or caregivers*. The five core symptoms of insomnia that we obtained (Figure 11) are *difficulty falling asleep, easy to wake up at night, dysphoria and irascibility, forgetful,* and *spiritlessness and weakness*. We further discovered the related symptoms corresponding to the core symptoms such as *irritability, dryness of mouth,* and *sweating at night*, which are all derived from the same syndrome. These findings also indicate the main syndrome for different individual cases. Therefore, our results essentially match the diagnostic criteria for the core symptoms of insomnia.

After training the node embeddings in the symptom network using the Skip-Gram model with different embedding dimensions (128 and 164), we acquired the different symptom embedding representations. We then clustered these symptom embeddings using the spectral clustering algorithm with different cluster numbers (4 and 5), and obtained four and five symptom communities, respectively. By comparing the experimental results with different dimensions and cluster numbers, we found that the clusters of insomnia symptoms are more identical to those in clinical practice and the results from previous studies when the dimension of the Skip-Gram model was 128 and the number of clusters in the spectral clustering algorithm was 5. Thus, the network shown in Figure 11 can reflect the distinct clinical symptom characteristics of insomnia, and each community is significantly heterogeneous, which will be helpful to evaluate the condition and guide individualized treatment.

Limitations

To best evaluate the results of core symptom identification or symptom clustering, we have simply presented the conclusion based on the symptom network structure analysis, evaluation metrics of node centrality in a complex network, and the similarity of symptom embeddings. The results were derived from objective calculations using machine-learning approaches. We also referred to the professional suggestions from clinicians working on insomnia, published manuscripts, and guideline for the diagnosis and treatment of insomnia. Because there is still no standard category for each symptom in TCM, the accuracy of the results remains to be verified.

Conclusions

In the clinical practice of TCM, the symptoms of insomnia patients with different syndromes are different. Therefore, research focused on the identification of core symptoms, syndromes, and their corresponding symptoms has significance for the clinical diagnosis and treatment of insomnia. By using complex network and machine-learning approaches, specifically

node-embedding and the spectral clustering algorithm, we constructed the symptom-weighted network model representing the relationships underlying the different symptoms. The insomnia symptoms were divided into five communities according to their distinct clinical characteristics. Multiple interrelated symptoms were frequently observed in the same community, reflecting the fact that different symptoms are derived from the same syndrome. These results can provide meaningful symptom associations, which can help physicians to find the most significant content and regularity from complex symptom relationships.

A similar diagnosis of symptoms appeared in a report by the Committee of the American Academy of Sleep Medicine [37]. Overall, the establishment of different communities can help to explore meaningful symptom associations, which can provide an intuitive understanding of the corresponding basic pathogenesis for physicians. Further, these results clarify that the methodologies used in this study can effectively and accurately find hidden relationships between symptoms for insomnia. These methodologies can filter unimportant symptoms and obtain meaningful symptom correlations and associations, which will help physicians to find the most important core content from complex symptom relationships. The trained insomnia symptom embeddings can be used in additional research as a basic dataset. With further development, similar approaches can be used to explore the symptom distribution regularity for the diagnosis and treatment of other diseases.

Acknowledgments

This study was funded by the National Natural Science Foundation of China (81874414), and the Natural Science Foundation of Hubei Province (2018CFB259).

Conflicts of Interest

None declared.

References

1. Shahin M, Ahmed B, Hamida ST, Mulaffer FL, Glos M, Penzel T. Deep Learning and Insomnia: Assisting Clinicians With Their Diagnosis. IEEE J Biomed Health Inform 2017 Nov;21(6):1546-1553. [doi: 10.1109/jbhi.2017.2650199]
2. Emert SE, Tutek J, Lichstein KL. Associations between sleep disturbances, personality, and trait emotional intelligence. Pers Individ Diff 2017 Mar;107:195-200. [doi: 10.1016/j.paid.2016.11.050]
3. 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 Jun;98(24):e15967 [FREE Full text] [doi: 10.1097/MD.0000000000015967] [Medline: 31192935]
4. Li F, Xu B, Wang P, Liu L. Traditional Chinese medicine non-pharmaceutical therapies for chronic adult insomnia. Medicine 2019;98(46):e17754. [doi: 10.1097/md.0000000000017754]
5. Allamanis M, Barr ET, Devanbu P, Sutton C. A Survey of Machine Learning for Big Code and Naturalness. ACM Comput Surv 2018 Sep 06;51(4):1-37. [doi: 10.1145/3212695]
6. Hu F, Liu J, Li LH, Liang J. Community detection in complex networks using Node2vec with spectral clustering. Physica A 2019 Nov 23:123633. [doi: 10.1016/j.physa.2019.123633]
7. Su Q, Zhu Y, Jia Y, Li P, Hu F, Xu X. Sedimentary Environment Analysis by Grain-Size Data Based on Mini Batch K-Means Algorithm. Geofluids 2018 Dec 02;2018:1-11. [doi: 10.1155/2018/8519695]
8. Hu F, Wang MZ, Zhu YH, Liu J, Jia YL. A time simulated annealing-back propagation algorithm and its application in disease prediction. Mod Phys Lett B 2018 Sep 05;32(25):1850303. [doi: 10.1142/s0217984918503037]
9. Zhou X, Peng Y, Liu B. Text mining for traditional Chinese medical knowledge discovery: a survey. J Biomed Inform 2010 Aug;43(4):650-660 [FREE Full text] [doi: 10.1016/j.jbi.2010.01.002] [Medline: 20074663]
10. Ahuja R, Vivek V, Chandna M, Virmani S, Banga A. Comparative study of various machine learning algorithms for prediction of insomnia. In: Chakraborty C, editor. Advanced Classification Techniques for Healthcare Analysis. Hershey, PA: IGI Global; 2019:234-257.
11. Park K, Lee S, Wang S, Kim S, Lee S, Cho S, et al. Sleep prediction algorithm based on machine learning technology. Sleep 2019;42:A172. [doi: 10.1093/sleep/zsz067.425]
12. Seixas A, Henclewood D, Langford A, McFarlane S, Zizi F, Jean-Louis G. Protective sleep and physical activity profiles in diabetes risk among blacks and whites in the United States: A Bayesian belief network machine learning model of national health interview survey. Sleep 2018;41:A324. [doi: 10.1093/sleep/zsy061.872]
13. Hu F, Qiao YL, Xie GJ, Zhu YH, Jia YL, Huang PP. Symptom distribution regulation of core symptoms in insomnia based on Infomap-SA algorithm. 2017 Oct Presented at: 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES); October 13-16, 2017; Anyang p. 229-232. [doi: 10.1109/DCABES.2017.57]
14. Hu F, Li LH, Huang XY, Huang PP, Chen L. On herb compatibility rule of insomnia based on machine learning approaches. 2019 Nov Presented at: 18th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES); November 8-10, 2019; Wuhan p. 257-260. [doi: 10.1109/DCABES48411.2019.00071]

15. Li LX, Liu Y, Wang N, Hou JA, Wang HS, Zhou ZX, et al. Study on pre-processing methods of clinical data from TCM individual treatment of insomnia based on syndrome differentiation. Chinese J Inf Tradit Chinese Med 2017;24(12):92-96. [doi: 10.3969/j.issn.1005-5304.2017.12.023]
16. Weng S, Zhou N. Analysis on zhong yi-tang's medication rule in prescriptions for insomnia based on data mining method. J Zhejiang Chinese Med Univ 2015;8:595-597. [doi: 10.16466/j.issn1005-5509.2015.08.006]
17. Shi C, Hu B, Zhao WX, Yu PS. Heterogeneous Information Network Embedding for Recommendation. IEEE Trans Knowl Data Eng 2019 Feb 1;31(2):357-370. [doi: 10.1109/tkde.2018.2833443]
18. Hamilton W, Ying R, Leskovec J. Representation learning on graphs: Methods and applications. arXiv 2017 Sep 17:1709.05584. [doi: 10.1093/oseo/instance.00178455]
19. Yang K, Wang N, Liu G, Wang R, Yu J, Zhang R, et al. Heterogeneous network embedding for identifying symptom candidate genes. J Am Med Inform Assoc 2018 Nov 01;25(11):1452-1459. [doi: 10.1093/jamia/ocy117] [Medline: 30357378]
20. Wang N, Li P, Hu X, Yang K, Peng Y, Zhu Q, et al. Herb Target Prediction Based on Representation Learning of Symptom related Heterogeneous Network. Comput Struct Biotechnol J 2019 Jan;17:282-290 [FREE Full text] [doi: 10.1016/j.csbj.2019.02.002] [Medline: 30867892]
21. Guo M, Yu Y, Wen T, Zhang X, Liu B, Zhang J, et al. Analysis of disease comorbidity patterns in a large-scale China population. BMC Med Genomics 2019 Dec 12;12(Suppl 12):177 [FREE Full text] [doi: 10.1186/s12920-019-0629-x] [Medline: 31829182]
22. Yang K, Wang R, Liu G, Shu Z, Wang N, Zhang R, et al. HerGePred: Heterogeneous Network Embedding Representation for Disease Gene Prediction. IEEE J Biomed Health Inform 2019 Jul;23(4):1805-1815. [doi: 10.1109/jbhi.2018.2870728]
23. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. 2014 Aug Presented at: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 22-27, 2014; New York p. 701-710. [doi: 10.1145/2623330.2623732]
24. Grover A, Leskovec J. Node2vec: Scalable feature learning for networks. 2016 Aug Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 22-27, 2016; New York p. 855-864. [doi: 10.1145/2939672.2939754]
25. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv 2013 Jan 16:1301.3781.
26. Newman MEJ. The Structure and Function of Complex Networks. SIAM Rev 2003 Jan;45(2):167-256. [doi: 10.1137/s003614450342480]
27. Hu F, Zhu YH, Liu J, Jia YL. Computing communities in complex networks using the Dirichlet processing Gaussian mixture model with spectral clustering. Phys Lett A 2019 Feb;383(9):813-824. [doi: 10.1016/j.physleta.2018.12.005]
28. Hu F, Liu Y. Multi-index algorithm of identifying important nodes in complex networks based on linear discriminant analysis. Mod Phys Lett B 2015 Feb 04;29(03):1450268. [doi: 10.1142/s0217984914502686]
29. Freeman LC. Centrality in social networks conceptual clarification. Soc Netw 1978 Jan;1(3):215-239. [doi: 10.1016/0378-8733(78)90021-7]
30. Bonacich P. Some unique properties of eigenvector centrality. Soc Netw 2007 Oct;29(4):555-564. [doi: 10.1016/j.socnet.2007.04.002]
31. Lin L. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 1989 Mar;45(1):255-268. [doi: 10.2307/2532051]
32. Nesselroade KPJ, Grimm LG. Statistical Applications For The Behavioral And Social Sciences. Hoboken, NJ: John Wiley & Sons Inc; 2018.
33. Fiedler M. Algebraic connectivity of graphs. Czechoslovak Math J 1973;23(2):298-305 [FREE Full text]
34. von Luxburg U. A tutorial on spectral clustering. Stat Comput 2007 Aug 22;17(4):395-416. [doi: 10.1007/s11222-007-9033-z]
35. Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 2006 Sep 11;74(3). [doi: 10.1103/physreve.74.036104]
36. Chesson A, Hartse K, McDowell W, Davila D, Johnson S, Littner M, et al. Practice parameters for the evaluation of chronic insomnia. Sleep 2000;23(2):237-242. [doi: 10.1093/sleep/23.2.1k]
37. Riemann D, Baglioni C, Bassetti C, Bjorvatn B, Dolenc Groselj L, Ellis JG, et al. European guideline for the diagnosis and treatment of insomnia. J Sleep Res 2017 Dec 05;26(6):675-700. [doi: 10.1111/jsr.12594] [Medline: 28875581]

Abbreviations

EMR: electronic medical record
TCM: traditional Chinese medicine

Edited by T Hao, Z Huang, B Tang; submitted 26.10.19; peer-reviewed by X Zhou, S Han; comments to author 12.01.20; revised version received 31.01.20; accepted 10.02.20; published 16.04.20

Please cite as:
Hu F, Li L, Huang X, Yan X, Huang P
Symptom Distribution Regularity of Insomnia: Network and Spectral Clustering Analysis
JMIR Med Inform 2020;8(4):e16749
URL: http://medinform.jmir.org/2020/4/e16749/
doi: 10.2196/16749
PMID: 32297869