Research on Traditional Chinese Medicine Domain Knowledge Graph Completion and Quality Evaluation.

Chang Liu, Zhan Li, Jian-Min Li, Yi-Qian Qu, Ying Chang, Qing Han, Ling-yong Cao, Shu-yuan Lin

Submitted to: JMIR Medical Informatics
on: December 02, 2023

Table of Contents

Research on Traditional Chinese Medicine Domain Knowledge Graph Completion and Quality Evaluation.

Chang Liu[1*]; Zhan Li[1*]; Jian-Min Li[2]; Yi-Qian Qu[1]; Ying Chang[1]; Qing Han[1]; Ling-yong Cao[3*]; Shu-yuan Lin[1*]

[1]School of Basic Medical Sciences Zhejiang Chinese Medical University Hangzhou CN
[2]Zhejiang Chinese Medical University and Gancao Doctor Chinese Medicine Artificial Intelligence Joint Engineering Center Zhejiang Chinese Medical University Hangzhou CN
[3]School of Basic Medical Sciences Zhejiang Chinese Medical University Hangzhou FR
[*]these authors contributed equally

Corresponding Author:
Shu-yuan Lin
School of Basic Medical Sciences
Zhejiang Chinese Medical University
School of Basic Medical Sciences, Zhejiang Chinese Medical University, 548 Binwen Road, Binjiang Dis
Hangzhou
CN

Abstract

Background: Knowledge graphs (KGs) can introduce domain knowledge into the traditional Chinese medicine (TCM) intelligent syndrome differentiation model. However, the construction quality of current KGs in the TCM field is uneven, which is related to lacking knowledge graph completion (KGC) and evaluation methods.

Objective: To explore KG completion methods and evaluation methods suitable for TCM domain knowledge.

Methods: In the KGC phase, according to TCM domain knowledge characteristics, we propose the three-step "entity- ontology-path" completion plan, using path reasoning, ontology rule reasoning and association rules. In the KGC quality evaluation phase, we propose a three-dimensional evaluation system of completeness, accuracy, and usability using quantitative indicators such as complex network analysis, ontology reasoning, and graph representation. Furthermore, we discuss the influence of different graph representation models on KG usability.

Results: In the KGC phase, 52, 107, 27, and 479 triples were added by outlier analysis, rule-based reasoning, association rules, and path-based reasoning, respectively. In addition, rule-based reasoning identified 14 contradictory triples. In the KGC quality evaluation phase, in terms of completeness, KG after completion had higher density and lower sparsity, and there were no contradictory rules in the KG. In terms of accuracy, KG after completion was more consistent with prior knowledge. In terms of usability, the MRR, MR, and Hist@N of the TransE, RotatE, DistMult, and ComplEx graph representation models all showed improvement after KG completion. Among them, the RotatE model achieved the best representation.

Conclusions: The three-step completion plan can effectively improve the completeness, accuracy and availability of KGs, and the three-dimensional evaluation system can be used for comprehensive KGC evaluation. In the TCM field, the RotatE model performed better in KG representation.

(JMIR Preprints 02/12/2023:55090)
DOI: https://doi.org/10.2196/preprints.55090

Preprint Settings

Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Chang Liu[1,2,3*], Zhan Li[1*], Jian-min Li[3], Yi-qian Qu[1,3], Ying Chang[1], Qing Han[1,3], Ling-yong Cao[1,3], Shu-yuan Lin[1,3**]

[1]School of Basic Medical Sciences, Zhejiang Chinese Medical University, 548 Binwen Road, Binjiang District, Hangzhou 310053, China.

[2]Traditional Chinese Medicine Hospital of Guangdong Province, 111 Dade Road, Yuexiu District, Guangzhou 510120, China.

[3] Zhejiang Chinese Medical University and Gancao Doctor Chinese Medicine Artificial Intelligence Joint Engineering Center, 548 Binwen Road, Binjiang District, Hangzhou 310053, China.

*Cofirst author
**Corresponding author
Prof Shuyuan Lin PhD
School of Basic Medical Sciences, Zhejiang Chinese Medical University, 548 Binwen Road, Binjiang District, Hangzhou 310053, China.
Phone: 86-18042323417
Email address: lin_shuyuan@foxmail.com.

E-mail addresses: acuwomen@126.com (C. Liu), 15294603632@163.com (Z. Li), 15600561914@163.com (J. Li), 94039363@qq.com (Y. Qu), changying228@163.com (Y. Chang), hanqing@zcmu.edu.cn (Q. Han), caolingyong@163.com (L. Cao), lin_shuyuan@foxmail.com (S. Lin).

Research on Traditional Chinese Medicine Domain Knowledge Graph Completion and Quality Evaluation

Abstract

Background: Knowledge graphs (KGs) can integrate domain knowledge into traditional Chinese medicine (TCM) intelligent syndrome differentiation model. However, the quality of current KGs in TCM domain varies greatly, which is related to the lack of knowledge graph completion (KGC) and evaluation methods.
Objective: This study aims to investigate KG completion methods and evaluation methods tailored for the TCM domain knowledge.
Methods: In the KGC phase, according to the characteristics of TCM domain knowledge, we propose a three-step "entity- ontology-path" completion approach. This approach employs path reasoning, ontology rule reasoning and association rules. In the KGC quality evaluation phase, we propose a three-dimensional evaluation framework that encompasses completeness, accuracy, and usability, utilizing quantitative metrics such as complex network analysis, ontology reasoning, and graph representation. Furthermore, we compare the impact of different graph representation models on KG usability.
Results: In the KGC phase, 52, 107, 27, and 479 triples were added by outlier analysis, rule-based reasoning, association rules, and path-based reasoning, respectively. In addition, rule-based reasoning identified 14 contradictory triples. In the KGC quality evaluation phase, in terms of completeness, KG after completion had higher density and lower sparsity, and there were no contradictory rules within the KG. In terms of accuracy, KG after completion was more consistent with prior knowledge. In terms of usability, the MRR, MR, and Hist@N of the TransE, RotatE, DistMult, and ComplEx graph representation models all showed improvement after KG completion. Among them, the RotatE model achieved the best representation.
Conclusions: The three-step completion approach can effectively improve the completeness, accuracy and availability of KGs, and the three-dimensional evaluationframework can be used for comprehensive KGC evaluation. In the TCM field, the RotatE model performed better in KG representation.

Keywords: graph completion; traditional Chinese medicine; graph quality evaluation; graph representation.

Introduction

Background

Traditional Chinese medicine (TCM) has unique advantages in diagnosing and treating a variety of diseases[1]. It also played a remarkable role in preventing and treating COVID-19 during the global pandemic[2]. The prerequisite to the effectiveness of TCM lies on accurate syndrome differentiation and treatment determination. However, the manual syndrome differentiation process has subjective differences[3]. Applying artificial intelligence technology to auxiliary diagnosis and treatment will contribute to standardization in this area[4]. A review has suggested that the accuracy of TCM intelligent syndrome differentiation models has reached the application standard[5]. However, deep learning models that perform well often suffer from a lack of explainability, and they are heavily reliant on data, which limits their application. Knowledge graphs (KGs) can integrate domain knowledge into intelligent models, reduce data dependency, and enhance explainability[6]. Therefore, many studies have constructed KGs in the TCM field. However, the variable quality of existing KGs[7] impacts their ability to effectively represent knowledge and support tasks such as intelligent diagnosis, question answering, and prescription recommendation. Given that the technology for constructing KGs is becoming increasingly pervasive, we contend that the absence of a comprehensive knowledge graph completion (KGC) and quality evaluation system tailored to TCM is a critical factor contributing to this variation.

Research on KGC and quality evaluation is essential in the field of TCM. Firstly, the existing construction work of TCM KGs predominantly relies on a single knowledge source and lacks a methodology for exploring rules from different knowledge sources. Secondly, there is a scarcity of methods to identify abnormal connections within KGs. KG completion necessitates a foundation of accurate knowledge. however, semiautomatic KG construction may incorporate contradictory, erroneous, or incomplete knowledge[8], which needs to be discovered and corrected. Currently, the TCM field KG lacks a systematic approach for identifying inaccurate knowledge and rectifying errors. Thirdly, the majority of existing methods for graph completion evaluation focus solely on specific algorithm evaluation metrics. lacking a stereoscopic evaluation framework that assesses the overall quality of the completed graph.

There are two current challenges: (1) Theories and methods for KGC are scarce within the TCM field, which makes it difficult to complete knowledge at different levels and identify inaccurate knowledge in the graph. (2) There is a lack of KGC quality management system and evaluation criteria that are specifically tailored for TCM domain knowledge.

Goal of This Study

To address these challenges, we have designed a completion plan based on the characteristics of TCM domain knowledge. This plan targets the three levels of knowledge—explicit, implicit, and tacit—and systematically completes the KG from the perspectives of path, ontology, and entity. We have also proposed a completion evaluation system that includes three dimensions: completeness, accuracy, usability, for each of these dimensions, we have developed specific evaluation metrics.

The contributions of this paper are summarized as follows. (i) Based on the characteristics of TCM knowledge, a three-step completion plan consisting of ‘path-ontology-entity’ is proposed. This plan not only identifies and corrects inaccurate knowledge but also enhances the completeness of the graph, providing a methodological reference for related research in the field. (ii) Under the KG quality management framework, we propose a quality evaluation system for TCM KGC, providing a reference for comprehensive and multidimensional KG evaluation and promoting KG quality improvement in the field.

Related Works

Application and Development of KG in TCM

The application of knowledge graph technology in the field of traditional Chinese medicine (TCM) dates back over 20 years. TCMLS (Traditional Chinese Medicine Language System) defines the most basic semantic types and semantic relationships in the field of TCM[9]. GFO-TCM is a mid-level ontology that is built upon the foundation of UTCMLS using a top-down approach[10]. Based on modern literature or by integrating multiple literatures, researchers have constructed knowledge graphs in various subdomains of TCM, including syndromes[11], medical cases[12,13], prescriptions and herbs[14], and health preservation[15], among others. KG has been applied in the TCM field in information retrieval, question answering [16-18], visual analysis[19], auxiliary diagnosis[20] and treatment, among others. However, there are still shortcomings in the explication of the syndrome differentiation process, the fusion of ancient and modern knowledge, and the combination of theory and clinical practice [21]. The study of KG should effectively address practical problems in TCM clinical practice and integrate the characteristics of the TCM knowledge system[22]. Therefore, this study focuses on the reports of KG in TCM auxiliary diagnosis and treatment. SUN et al. built a TCM auxiliary diagnosis and treatment system for rheumatoid arthritis, which was based on the knowledge from TCM classics, providing doctors with guidance on diagnosis and treatment knowledge [23]. FU et al. constructed a knowledge graph of acute abdomen usingNeo4j, employed a diagnosis and treatment reasoning algorithm based on association rule mining combined with random walk, and provided information services and technical support for primary doctors by recommending personalized diagnosis and treatment plans for cases [24].

In current research, the intelligent syndrome differentiation in TCM is frequently represented as a classification problem, where deep learning models receive symptom information as input and output syndrome categories. Graph-based representation learning can provide domain knowledge to the model. For instance, LI transformed 20,000 medical records into medical record graphs and utilized them as inputs to graph convolutional networks to learn graph embedding of prostate cancer features[25]. This approach effectively maps the features of prostate cancer and facilitates the diagnostic process. LI embedded knowledge regarding cerebral palsy from KGs into tensors and integrated them into recurrent neural networks, achieving a diagnostic accuracy of 79.31%. Subsequent fine-tuning with electronic medical records elevated the model's accuracy to 83.12%[26].

Overview of KGC Methods in the Medical Field

KGC is an application of knowledge reasoning. Knowledge reasoning is essential for addressing the incompleteness, potential biases and errors found in KGs, as well as for inferring hidden information between knowledge entities[27]. Methods for KGC can be broadly categorized into three types: rule-based, vector-based, and neural network-based.

Rule-based reasoning[28] has the advantage of utilizing prior knowledge to provide accurate and traceable reasoning with high explainability. However, the downside is the difficulty in enumerating all the rules, and the limited generalization ability. Rule-based reasoning includes predicate logic rules and ontology rules.

Vector-based reasoning methods first project entities and relations into a vector space. Triplets serve as input to learn vector representations through constraint functions. Predicted triplets are generated by fixing the head entity and applying a representation model. By converting reasoning problems into vector calculation problems, vector-based reasoning is more efficient and easier to train than traditional reasoning approaches. Common graph representation learning models include TransE[29], TransH[30], TransR[31], TransD[32], RotatE[33], RESCAL[34], DistMult[35] and ComplEx[36]. However, this approach primarily focuses on direct relations between entities and overlooks indirect paths among entities in graphs. In addition, it lacks explainability. Moreover, embedding methods degrade with increasing sparsity and unreliability of the KG[37].

Neural network-based reasoning involves learning entity features and semantic sequences from prior knowledge and then utilizing neural networks to identify the linkage path between two entities to aid in reasoning by predicting the relation path. This approach has the advantage of utilizing the graph structure and hidden node information to the fullest extent possible. However, it also has drawbacks, including high model complexity, large data requirements, and poor explainability.

Current Status of KG Quality Assessment

XUE[38] summarized the current research on KG quality management and proposed five dimensions for KG evaluation: accuracy, consistency, completeness, timeliness, and redundancy.

The methods for KG quality evaluation can be categorized into four types: human-based, statistical-based, rule-based, and comprehensive. We reviewed the literature related to the completion and quality assessment of medical KGs over the past five years and classified the evaluation methods into the following three dimensions, with reference to XUE's definition:

(i) Completeness: A scale was designed, and medical experts were invited to manually evaluate KG data authority and data volume[39].

(ii) Accuracy: Using mean average precision (%MAP), an evaluation index for target detection and classification, the prediction triplet was evaluated as a classification problem[40]. Weighted sampling was conducted on the completed (predicted) triples. Experts judged the correctness of these triples, and the accuracy was calculated[41]. Through complex network analysis methods such as clustering[42] and t-SNE visualization[43] (a visualization method of data after dimensionality reduction), KG disease classification knowledge was summarized and compared with prior knowledge, aiming to determine whether the data distribution of the constructed graph was consistent with prior knowledge.

(iii) Usability: KG quality is evaluated by the effectiveness of graph representation. For instance, KG can be vectorized through graph representation algorithms to predict tail entities based on head entities and relationships. This is frequently used to compare the completion effects between different algorithms. Common metrics are the mean rank (MR) of the correct answer, the reciprocal rank of the first correct answer (MRR), and the normalized discount cumulative return of the first N predicted tail entities (Hist@N) [44]. Additionally, there are studies to evaluate KGs by examining their performance on downstream tasks, such as the introduction of area under the receiver operating characteristic curve indicators in drug reuse and target identification[45].

Among the above methods and indicators, the direct manual evaluation of KG was affected by subjective factors and was not adopted in this study. The %MAP metric is appropriate for predicting multi-category triples, which is not consist with the design of our study protocol. The purpose of this study is to design a general evaluation method for KG completion in the field of TCM. Since indicators such as receiver operating characteristic curve require specific downstream tasks for their application, the intermediate stage of KG utilization—KG representation—was chosen as the criterion for usability evaluation. With the remaining methods as reference, we designed a TCM KGC evaluation system (detailed in Section 3.2.3). To quantitatively assess the quality of KG, we introduce some metrics derived from complex network analysis.

Introduction of the TCM knowledge system

In TCM theory, *syndrome* (Zheng Hou) refers to the classification and summary of relatively stable symptoms and signs during disease occurrence and development. *Syndrome* is the diagnostic conclusion in TCM. For the sake of convenience, we will refer to the symptoms and signs that patients have as *symptoms* in the following text. The cognitive process of determining the syndrome is referred as *syndrome differentiation* (Bian Zheng). This process involves inferring the *pathogenesis factors* (Bing Ji, including disease location, disease nature, and disease state) from the symptoms and then composing the symptoms based on specific combinations and weights of the pathogenesis factors[46]. The process of determining the treatment plan is called *prescription determination* (Lun Zhi), which involves formulating the main prescription based on the syndrome and adjusting the medication according to the symptoms.

Among many syndrome differentiation methods, *The Six Channel Syndrome Differentiation* is one, which categorizes disease syndromes into six major categories: *Tai Yang, Yang Ming, Shao Yang, Tai Yin, Shao Yin*, and *Jue Yin*, and further subdivides each syndrome level under these major categories. For instance, *Tai Yang* includes three syndromes: *Tai Yang Shang Han*, *Tai Yang Zhong Feng*, and *Feng Han Liang Gan*. When a patient presents with *symptoms* such as aversion to cold, spontaneous sweating, and slow pulse, which indicate pathogenesis factors of external cold, deficient defense and weakened nutrients, the syndrome can be identified as *Tai Yang Zhong Feng,* and the main prescription would be *Cassia Twig Decoction* (Gui Zhi Tang) (as shown in Figure 1).

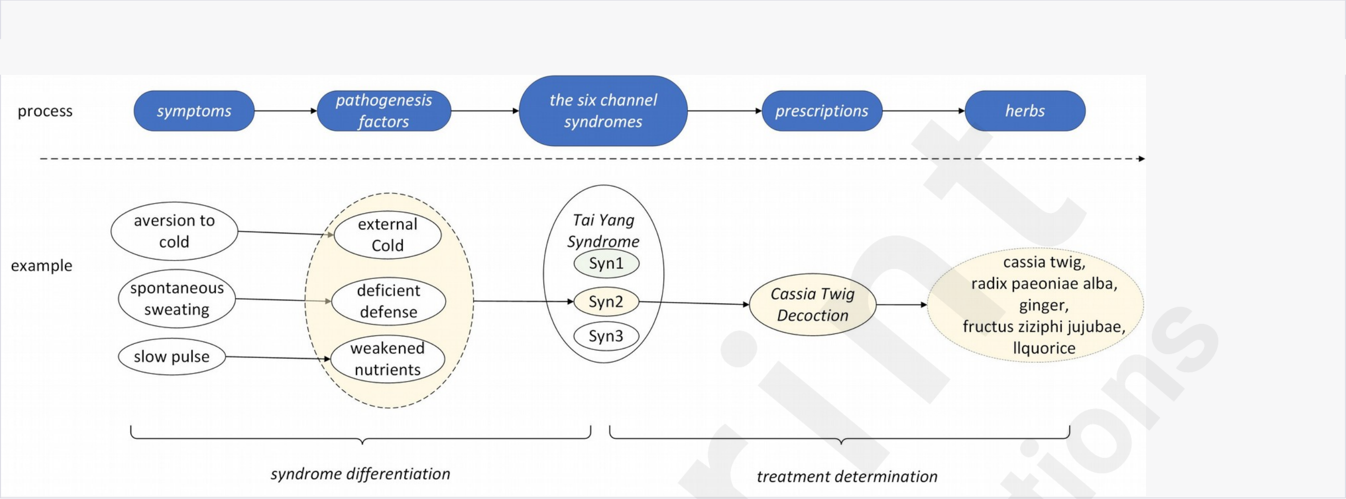


Fig. 1. *The Six Channel Syndrome Differentiation process*. Cassia Twig Decoction: □□□, Cinnamomi Ramulus: □□, Paconiae Radix Alba: □□, Zingiberis Rhizoma: □□, Ziziphus jujuba Mill: □□, Glycyrrhizae Radix Et Rhizoma: □□.

Referring to research in cognitive science on brain cognition and memory, human knowledge can be divided into three levels according to the different types of performance: explicit knowledge, implicit knowledge and tacit knowledge[47]. Its definition and embodiment in the TCM field are as follows:

(i) Explicit knowledge refers to the knowledge already present in the text, some of which can be derived from the first-order predicate logic of the original and others through the reasoning of multistep connections. Examples of this include the law of syndrome differentiation and the treatment rules in ancient medical books.

(ii) Implicit knowledge refers to knowledge that is not present in the text but exists in the domain scheme defined by the ontology. For example, the symptoms and contraindications that are not documented in ancient books can be inferred from explicit knowledge.

(iii) Tacit knowledge is not explicitly stated in the text and can be uncovered through data mining methods. For example, the clinical manifestations of comorbidities and syndromes that are present in Electronic Medical Records.

The KG completion plan in the TCM field should be designed based on the characteristics of knowledge at the above three levels.

Discussion on the KG completion method in the TCM field

The medical field requires knowledge that is accurate, rigorous and traceable , hence KG completion should be interpretable. This requirement influences our selection of completion methods. Although graph representation learning is relatively popular, it was employed solely as an evaluation method in this study due to its limited explainability.

Path inference can fully utilize the paths between nodes for rule mining. The classic PRA algorithm learns KG relation characteristics through random walks and can predict potential relationships between two entities using the path between them[48]. Path-based reasoning has good performance and explainability[49]. LIU optimized the link prediction model based on the path ranking algorithm, which was applied for TCM KG completion of famous prescriptions[50]. SHAO et al. constructed KGs for famous TCM doctor experiences in diagnosing and treating lung cancer, employed the RED-GNN model, and mined implicit knowledge using relational path reasoning[51]. Outlier detection[52] can be considered as a special form of path reasoning that can automatically identifies abnormal connections in the KG. This study introduces outlier detection to improve KG accuracy.

Data mining methods have been widely used to discover hidden rules in TCM[53]. Association rule mining, a prevalent data mining method in TCM, explores the relation between item sets in datasets. It is frequently applied to mine relations between symptoms and syndromes, medications and syndromes, as well as symptoms and medications[54]. The Apriori algorithm is a Boolean, single-dimensional, and single-layer association rule that links and prunes all item sets generated by multiple scans, leveraging the Apriori property to improve mining efficiency[55]. During the process of syndrome differentiation, core information is derived from knowledge associated with symptoms, pathogenesis factors, and TCM syndromes. The relation between 'symptoms' and 'pathogenesis factors' is relatively consistent. Consequently, the primary focus of KGC is to mine rules between 'symptoms' and 'syndromes'. Additionally, the (prescription-treat-symptom) triples can provide supplementary information to the 'symptoms' vector in KGE. Therefore, KGC should primarily focus on completing the above two types of relations. Although these relations may be frequently absent or irregularly distributed in ancient literature, they are readily available in clinical case data. For instance, ancient literature lacks records of tongue and pulse manifestations for diabetes, which is also known in TCM as 'Xiao Ke'.

In summary, path reasoning is suitable for reasoning explicit knowledge, ontology-based reasoning is suitable for mining implicit knowledge, and data mining is suitable for discovering tacit knowledge.

Methods

TCM KG Completion Methodology

Based on the characteristics of TCM knowledge and targeting the three-level knowledge system of "explicit knowledge, implicit knowledge and tacit knowledge", this study constructs a three-order completion plan of "path-ontology-entity":

First Order -- Complete explicit knowledge at the "path" level. Focus on the unique structure of the path in the KG, and mine knowledge using multistep predicate logic based on path reasoning.

Second Order -- Complete the "ontology" level of implicit knowledge. Employ ontology-based rule reasoning to make implicit framed knowledge explicit by generating new triples.

Third Order -- Complete tacit knowledge at the "entity" level. Utilize data mining methods to identify unestablished associations between entities. This paper takes association rule mining, a widely employed data mining method, as an example.

Proposed Method

Task description

We focuses on the completion and evaluation of TCM domain KGs, aiming to achieve a completion that improves both accuracy and completeness and to comprehensively evaluate the quality of the graph after completion. Specifically, the task includes the following steps: (1) KG completion. (i) Explicit knowledge completion: Identify the isolated triples in the recognition graph by detecting outliers and mine(Syndrome-Manifest-Symptom) and (Prescription-Treat-Symptom) rules from the clinical case dataset employing path-based reasoning. (ii) Implicit knowledge completion: Use ontology-based deductive reasoning and the discovery of contradictory knowledge to supplement missing knowledge and correct inaccuracies. (iii) Tacit knowledge completion: Employ association rules to mine (Syndrome-Manifest - Symptom) knowledge in KG. These generated triples are incorporated into the graph. (2) KGC evaluation encompasses three dimensions: completeness, accuracy, and usability. Metrics specific to each dimension are employed to compare the graph before and after completion, thereby assessing the quality of completion. Specifically, completeness evaluation is grounded in statistical methods, which characterize the graph by complex network features. Accuracy evaluation employs a multifaceted approach, incorporating ontology reasoning and complex network centrality analysis. Usability evaluation is based on a statistical method that compares the impact of KGC with that of KGE. The methodology is shown in Figure 3.

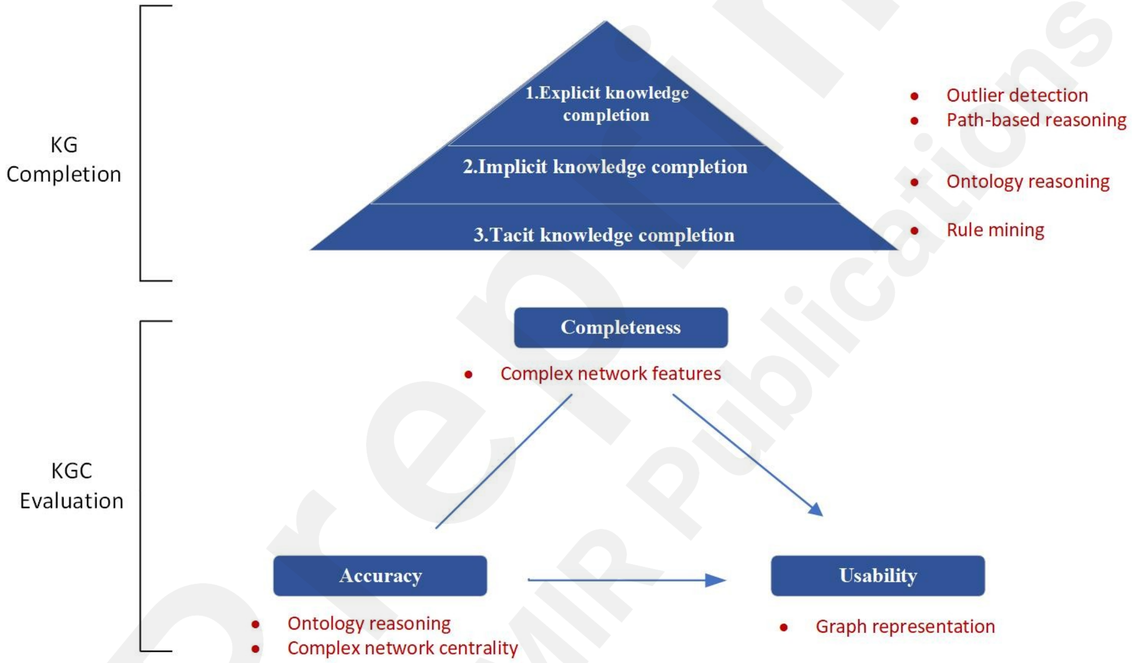


Fig. 3. Methodology of research.

KG completion

(i) Explicit knowledge completion

(a) Outlier detection: Outliers are data points that deviate from the norm of the dataset and are deemed inconsistent with the rest of the data. In our study, outliers are defined as isolated subgraphs that lack interconnections with other subgraphs. These isolated subgraphs were manually reviewed and categorized, and ontology reasoning was utilized to complete the relations among subgraphs that provide valuable diagnostic information, thereby integrating them with other subgraphs.

(b) Relation prediction based on path inference: The path ranking algorithm was used to mine potential relations involving the antecedents of 'syndrome' or 'prescription' and the consequent of 'symptom'. The results were ranked based on the number of supporting paths. Feature extraction involved generating paths using a random walk approach and selecting the feature set of paths. Feature calculation entailed computing the feature value $P(s \rightarrow t; \pi_j)$ for each training sample, which represents the probability of transitioning from node s to node t through relationship path π_j. A classifier was trained using the feature values of the training samples and utilized to infer the existence of a target relationship between two entities. The score function is:

$$score(s,t)=\sum_{\pi j\in pl}\theta_j P(s\rightarrow t;\pi_j) \quad (1)$$

Rules with more than two supporting paths were selected. The predicted results were further filtered using ontology inference based on the triples of (Symptom-Correspond to-Pathogenesis Factors) and (Prescription-Treat-Syndrome). The 'symptom' in the rules was converted to 'pathogenesis factor', and the 'prescription' was converted to 'curable syndrome'. Only the rules that consistent with the triples of (Syndrome-Contains-Pathogenesis Factors) in the KG were selected. The antecedent and consequent of the rules were used as the head and tail entities of the predicted triples, respectively. These entities were connected by relation to obtain two types of predicted triples: (Syndrome-Manifest-Symptom) and (Prescription-Treat-Symptom).

The accuracy, recall, and F1-score of the predicted triples were calculated using back-to-back annotations from two experts as the standard. The predicted triples from both methods were merged into the KG for completion.

$$\text{Precision}=TP/(TP+FP)$$
$$\text{Recall}=TP/(TP+FN)$$
$$F1=(2P\cdot R)/(P+R) \quad (2)$$

(ii) Implicit knowledge completion

Ontology reasoning: Based on the description logic of the ontology, the triples within the KG were completed and corrected. The relation properties, mainly involving transitivity, symmetry, and mutual exclusivity, were defined as shown in Table 1. Triples featuringtransitive and symmetric relations were reasoned, and the deduced triples were integrated into the graph. Contradictory triples were detected through mutual exclusivity and corrected upon expert review.

Table 1. Definition of relations' properties in ontology

Property	Definition	Relation
Transitivity	Relation P to ∀ entity x, y, z: P (x, y) and P (y, z) include P (x, y),	Contain (for relation 'clinical manifestation is')
Symmetry	Relation P to ∀ entity x, y: P (x, y) equal to P (y, z),	Differential Diagnosis Is
Mutual Exclusivity	Entity x, y simultaneously exists with P (x, y) and R (x, y), but a contradiction arises between relations P and R.	Treat and Contraindication Is

(iii) Tacit knowledge completion

Relationship prediction based on association rule mining: The Apriori algorithm was used to mine rules in medical records, with 'syndrome' as the antecedent and 'symptom' as the consequent. The reliability of the rules was evaluated by the value of lift, where a lift greater than 1 indicates a positive correlation between the two items. The support of item set X is defined as the proportion of transactions in the dataset that contain the item set. The confidence of a rule is defined as confidence(X=>Y), which can be interpreted as an estimate of the probability P(Y|X)[56]. The lift measure for a rule (X=>Y) is calculated as follows:

lift(X=>Y) = confidence(X=>Y)/support(Y) = P(Y|X)/P(Y). (3)

The parameters for the Apriori algorithm were set to a minimum support of 10%. The resulting rules were sorted by lift value, and rules with a lift greater than 1 were considered as the mining results for the model.

Quality evaluation

In this study, we proposed a KG completion evaluation system tailored for the TCM domain KG, which consists of three dimensions. (i) Completeness, which assesses whether the graph includes relevant data of interest in the domain. (ii) Accuracy, which measures the graph's reflection of facts, ensuring consistency with prior knowledge. (iii) Usability, which evaluates the difference in KGE before and after graph completion. Since the data quality dimension is abstract, specific measurement criteria were defined in this study to apply and quantify these dimensions in practice.

(i) Completeness

The overall structural completeness of the graph before and after completion was evaluated through topological properties. The specific indicators included the number of nodes, relations, degree, degree distribution (maximum degree, average degree), and network density. The degree of a node, denoted as k , was defined as the number of edges directly connected to a node. The average degree of all nodes in the network was denoted as <k>. The density ρ of a network with N nodes was defined as the ratio of the actual number of edges M to the maximum possible number of edges.

$$\rho = \frac{M}{\frac{1}{2}N(N-1)} \quad (4)$$

Network density can measure the sparsity of a network. A network density approaching zero indicates that the actual number of edges in the network is of lower order than N^2, and thus, the network is considered sparse.

(ii) Accuracy

Step 1: Integrated approach of Rule-based and human-based: We employed mutual exclusion rules in ontology reasoning to assess the alternations in contradictory knowledge before and after completion. (The method used was similar to that described in Section 3.2.2 (i).)

Step 2: Statistical method-based approach: Complex network centrality: We analyzed the distribution of symptoms and prescriptions before and after completion, and compared them with prior knowledge. The specific indicators and their meanings were identified as follows.

(a) Closeness centrality (CC): This metric quantifies how closely a node is to the center of the network. By calculating the CC for prescription nodes, we were able to identify the central prescriptions in the graph and compared them against prior knowledge. Note: In this study, central prescriptions as defined in prior knowledge were those primary and secondary prescriptions for treating syndromes listed in the 'Treatise on Febrile Diseases' textbook [57]. The CC calculation method for node i is:

$$CC_i = \frac{N}{\sum_{j=1}^{N} d_{ij}} \quad (5)$$

where dij represents the distance from node i to vertex j, and N represents the number of nodes in the network.

(b) K-core value: This refers to the maximum subgraph of a graph in which each node has a degree of at least k, and no more nodes can be added without reducing any node's degree below k. By using the k-core value of symptom nodes, we were able to pinpoint the symptom groups emphasized in the diagnostic system within the KG.

(iii) Usability

The effectiveness of graph completion was assessed through KG representation. We used all triples in the KG as samples and randomly divided them into training set and validation set in a 7:3 ratio. Various representation models were used to represent entities and relations. Negative samples were generated by replacing the tail entity of the actual triples with a randomly selected entity. Both positive and negative samples were fed into the model. During the training process, each example was assigned a loss function (see Table 2) to ensure that the score discrepancy between positive and negative samples exceeded the predefined margin, facilitating feedback for model updating. After each epoch of model training, the validation set was used to predict model performance. The L2 norm was adopted to measure the distance between the head entity's mapping vector and the tail entity, resulting in the predicted score for each entity when acting as a tail. By sorting all triples according to the predicted score of the tail entit[redacted] ranking of true triples among all triples can be obtained.

We mean rank (MR), mean reciprocal rank (MRR), and the proportion of correctl p N triples (Hist@N) before and after completion using the same model.

Table 2. Loss function of KG representation models

Model	Score Function
TransE	$-\|\|h+r-t\|\|$
DistMult	$\langle h,r,t \rangle$
ComplEx	$\mathrm{Re}(\langle h+r-t \rangle)$
RotatE	$-\|\|h \circ r-t\|\|$

$$MR = \frac{1}{|S|}\sum_{i=1}^{|S|} rank_i \qquad (6)$$

$$MRR = \frac{1}{|S|}\sum_{i=1}^{|S|} \frac{1}{rank_i} \qquad (7)$$

$$HITS@n = \frac{1}{|S|}\sum_{i=1}^{|S|} \mathbb{I}\,(rank_i \leq n) \qquad (8)$$

S is the set of triples, |S| is the number of triple sets, $rank_i$ is the prediction rank of the ith triple, and $\mathbb{I}$ is the indicator function (1 if the condition is true, 0 otherwise).

Results

Data sources

In our previous work, we developed a KG based on the ancient medical texts of *Treatise on Febrile and Miscellaneous Diseases.* We standardized and systematically organized the core concepts and their relations, particularly the *Six Channel Syndromes*, pathogenesis factors, and symptoms, through data mining and literature research. The graph comprises 1,255 nodes and 4,519 edges. Nodes and edges related to *Jue yin* syndrome are shown in Figure 2. Additionally, we standardized the medical records of 470 clinical cases treated with classical prescriptions for rule mining in KGC.

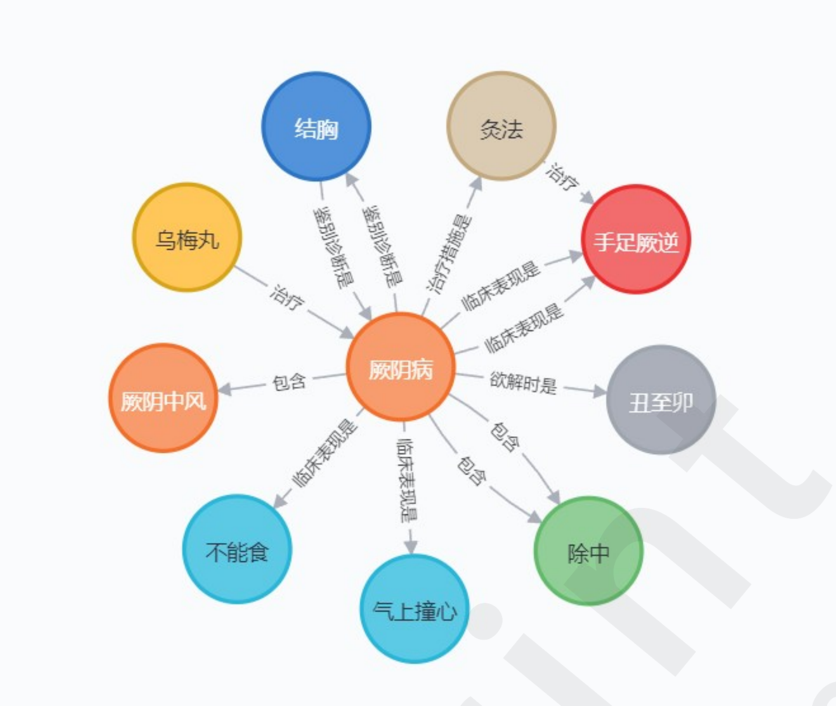


Fig. 2. Part of KG based on *Treatise on Febrile and Miscellaneous Diseases*. Node names: □□□: JueYin Syndrome, □□□□: JueYin Zhong Feng, □□□: Fructus Mume Pill, □□: Chest Binding, □□: moxibustion, □□□□: extremely cold limbs, □□□: 1 to 5 a.m., □□: Chu Zhong, □□□: no appetite, □□□□: a feeling of gas rushing up toward the thorax. Relation names: □□□□□: the clinical manifestation is, □□: include, □□: treat, □□□□□: differential diagnosis is, □□□□□: the treatment method is, □□□□: time for recovery.

KG Completion Results

Explicit knowledge completion

(a) Outliers: A total of 9 isolated subgraphs were identified, of which 6 were related to 'Fangzheng' (as shown in Table 3). 'Fangzheng' belongs to the subsyndromes under the secondary syndrome, which specifically refers to the syndrome treated by a certain prescription and is mainly used to represent knowledge related to differential diagnosis in the KG. By supplementing the (Prescription-Treat-Fangzheng) triple, connections between 'Fangzheng' and other subgraphs can be established. A total of 52 triples were added. The remaining 3 isolated subgraphs were related to treatment methods and seldom-used prescriptions, which were not completed.

Table 3. List of Outliers

Isolated Subgraphs	Nodes of Subgraphs	Related to
1	(Li Zhong Pill Syndrome)□□□□, (Red Halloysitum Rubrum and Limonitum Decoction Syndrome)□□□□□□□□, (Meretrix Powder Syndrome)□□□□, (Poria-Liquorice Decoction Syndrome)□□□□□□, (Wu Ling Powder Syndrome)□□□□, (Xie Xin Decoction Syndrome)□□□□, (Sanwubai Powder Syndrome)□□□□□	Fangzheng
2	(Capejasmine and Fermented Soybean Decoction	Fangzheng

	Syndrome)□□□□□, (Capejasmine-Ginger-Fermented Soybean Decoction Syndrome)□□□□□□□, (Capejasmine-Liquorice-Fermented Soybean Decoction Syndrome)□□□□□□□	
3	(Cassia Twif and Radix Aconiti Lateralis Preparata Decoction Syndrome)□□□□□□, (Cassia Twif and Radix Aconiti Lateralis Preparata plus Atractylodes Decoction Syndrome)□□□□□□□□□□□	Fangzheng
4	(No interior Syndrome)□□□, (Ephedra-Radix Aconiti Lateralis Preparata-Liquorice Decoction Syndrome)□□□□□□□□	Fangzheng
5	(Platycodon Grandiflorus Decoction Syndrome)□□□□, (Liquorice Decoction Syndrome)□□□□	Fangzheng
6	(Bulbus Allii Fistulosi and Sus Scrofa Domestica Brisson Decoction Syndrome)□□□□□□□□, (Bulbus Allii Fistulosi Decoction Syndrome)□□□□	Fangzheng
7	(heat pathogen)□, (curable)□□	treatment methods
8	(Fructus Terminaliae Chebulae)□□□, (porridge)□, (Frctus Terminaliae Chebulae Powder)□□□□	seldom-used prescriptions
9	(Sores of immersion)□□□, (Coptidis Rhizoma Powder)□□□	seldom-used prescriptions

(b) Path-based reasoning

A total of 1335 rules were mined, and 479 rules were selected through ontology reasoning. The comparison of the manual audit results and the model results showed an accuracy rate of 0.6124, a recall rate of 0.4906, and an F1 value of 0.5448. The observation reveals that path-based reasoning can extract a substantial number of potential rules; however, its accuracy is suboptimal. Employing ontology reasoning for further screening not only enhances the F1 score but also alleviates the burden of manual review.

Implicit knowledge completion

A total of 107 triples were added based on transitivity and symmetry, and 14 contradictory triples were discovered and removed based on mutual exclusion. Ontology-based reasoning can effectively identify inaccurate knowledge in KG.

Tacit knowledge completion

Using association rule mining, a total of 27 rules with lift values greater than 1 were discovered. Of these, 21 were related to *Yang Ming Tai Yin Combined Syndrome*, 4 to *Yang Ming Syndrome,* and 2 to *Jue Yin Syndrome.* The *Yang Ming Tai Yin combined syndrome* exhibits the highest number of associated rules, which can be attributed to its prevalence in medical cases. Rules with higherrepetition are easily mined by association rules.

All of the above rules were converted into triples and added to the graph.

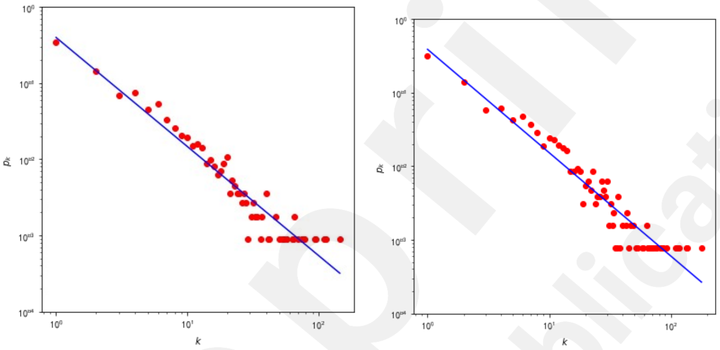
Quality Evaluation

KG after completion had higher density and lower sparsity, without contradictory rules. It was more consistent with prior knowledge, and improved the representation results of graph representation models.

Completeness

Table 4. Description of KG

	node	relation	largest k	average k	network density	slope of k distribution curve
before KGC	1255	4519	145	7.2016	0.0057	-1.4550
after KGC	1277	5162	179	8.0846	0.0063	-1.4058



before KGC after KGC

Fig. 4. Distribution curve of the KG.

The degree values both before and after KG completion followed a power-law distribution (see Figure 4), indicating a core KG structure. The core syndromes before and after completion were both primary syndromes of Six Channel Syndrome, which is consistent with prior knowledge. Since the completion mainly focused on relations, the network density increased and the sparseness decreased after completion (see Table 4).

Accuracy

The number of contradictory rules was 14 before completion, and it decreased to 0 after completion.

Among the top 20 prescriptions ranked by CC after completion, core prescriptions accounted for 80% in the KG after completion, which is a 5% increase from the KG before completion. This indicates that the completion work made the graph more consistent with domain knowledge.

Table 5. k-core value of symptom nodes before and after completion

	largest k-core	symptoms with largest k-core	proportion of symptoms related to exterior syndrome
before graph completion	10	32	0.3438
after graph completion	12	43	0.3721

The proportion of symptoms related to *Exterior Syndrome* among the symptoms with the highest k-core values increased after KGC increased (see Table 5). These symptoms include aversion to cold, floating pulse, fever, spontaneous sweating, aversion to wind, body pain, tight pulse, headache, anhidrosis, cold limbs, and chest tightness. Given that *Six Channel Syndrome Differentiation* emphasizes the differentiation of *Exterior Syndrome*, the completed graph is closer to prior knowledge.

Usability

Table 6. KGE performance before and after completion

	MRR	MR	H@1	H@3	H@10
Before completion					
TransE	0.2245	126.6173	0.1385	0.2502	0.3866
RotatE	0.3682	125.0212	0.3077	0.3734	0.4878
DistMult	0.2908	255.7703	0.2472	0.2991	0.3721
ComplEx	0.3196	244.0631	0.2835	0.3189	0.3913
After completion					
TransE	0.2414	114.0714	0.1507	0.2657	0.4256
RotatE	0.3830	115.0664	0.3130	0.4009	0.5281
DistMult	0.2944	233.3249	0.2512	0.3020	0.3731
ComplEx	0.3265	231.3083	0.2875	0.3235	0.4081

After completion, compared with before completion, the MR of each model decreases, the MRR is closer to 1, and Hist@N is increased, suggesting that the representation performance of each model is improved. Among them, the RotatE model changes the most. (see Table 6).

Discussion

Principal Results

We summarize the characteristics of TCM domain knowledge and design a three-step "path-ontology-entity" KG completion plan. The plan can efficiently complete explicit knowledge, effectively reason about implicit knowledge, and mine tacit knowledge, while maintaining good explainability. This paper exploresthe transfer of KG quality management systems to the TCM field and designs a comprehensive evaluation system for KGs in this field. The scheme is comprehensively evaluated from the three dimensions: completeness, accuracy and usability, each with its own set of quantitative indicators.

For the KG constructed around 'syndrome', core 'syndrome' nodes that establish more connections with other nodes can offer additional information for syndrome differentiation. When there is a discrepancy between core symptoms or core prescriptions in prior knowledge and the KG, it can be inferred that the KG has not fully represented the knowledge, which can guide researchers in subsequently completing the relationships of specific categories. Nodes with a higher k-core value are key points connecting other nodes in the KG and often provide differential diagnostic information. The KG completed and evaluated using the aforementioned methods, will provide accurate domain knowledge for tasks such as clinical decision support on syndrome differentiation and prescription recommendation.

We also explored KGE methods tailored for TCM KG. Our model with RotatE achieved the best performance, followed by ComplEx, while TransE performed the least effectively. TransE was unable to handle complex relationships such as one-to-many, many-to-one, and many-to-many. RotatE more effectively captured directional relationships between entities and handle complex graph structures, which aligns better with the characteristics of KGs in the TCM domain. In ComplEx, entity and relationship embeddings no longer exist in the real space but in the complex space, capturing more information. This study can provide a reference for other intelligent diagnosis and treatment research with KG fusion.

Comparison with Prior Work

Compared with the existing research, this study analyzes the characteristics of TCM domain knowledge and proposes a methodological theory for KG completion in the TCM domain, which enhance the systematic and comprehensive nature of the completion process. In addition, outlier detection is a completion method not used in existing studies. In terms of improving KG accuracy, this method can effectively identify missing knowledge in KGs, while ontology-based reasoning is more appropriate for identifying inconsistent knowledge. These two methods complement each other. While complex networks have been used to mine knowledge within KGs[58], there is a lack of literature on their use for graph quality assessment. The KG completion and evaluation approach proposed in this study for TCM domain may serve as a reference for KG construction in this field.

Limitations of the study: The KG constructed in this study is of a small scale. A thorough validation of the proposed methods is necessary when applied to larger or more diverse datasets. The methods employed in this paper do not occupy a lot of computational resources. However, the use of random walk approach may have higher time complexity than dynamic programming or heuristic search algorithms. Association rule mining extracted only a small amount of tacit knowledge, which may be related to the number and representativeness of medical records. In the medical cases, 45% of the syndromes were JueYin Syndrome, and 35% were YangMing - TaiYin Syndrome. The insufficient data quantity for other syndromes resulted in low lift values in the association rule mining and preventing the discovery of ‘Syndrome-Symptom’ associations. In subsequent studies, we plan to increase the number of medical records and explore other rule mining methods. For instance, generative adversarial networks can be used for data enhancement, thereby making the sample distribution more balanced.

Conclusions

The lack of KG completion and evaluation methodology restricts the development of KGs in the TCM field. This study first analyzes the knowledge levels within the TCM domain and proposes a three-step completion plan of "path-ontology-entity" based on the characteristics of each knowledge level: path reasoning is employed to mine explicit knowledge, ontology reasoning to mine implicit knowledge, and association rule analysis to mine tacit knowledge. An evaluation system including three dimensions—completeness, accuracy and usability—is designed, with each dimension using quantitative evaluation indicators to assess the quality of the completed KG. The results indicate that, under the guidance of the proposed methodology, the completed graph exhibits improvements across all dimensions. In terms of completeness, 22 nodes and 643 edges are added to the completed graph, and the network density is increased. In terms of accuracy, the core prescriptions among the top 20 CC prescriptions of KG after completion increased by 5% compared to those before completion, and the proportion of symptoms related to syndromes with the highest k-core value increased, suggesting that KG after completion is more in line with prior knowledge. In terms of usability, in the triplet prediction task, the completed KG enhances the performance of all graph representation models. The "path-ontology-entity" three-step completion plan effectively improve the integrity, accuracy and availability of KGs, and the three-dimensional evaluation system provides a comprehensive assessment of KGC. In addition, it was found that the RotatE modeloutperforms other commonly used models in the graph representation of KGs within the TCM domain. Our study provides a methodological reference for the completion and evaluation of TCM KGs.

Acknowledgements

Authors' contributions

Zhan Li, Chang Liu, and Shuyuan Lin proposed the concept and framework of this paper. Zhan li and Yiqian Qu constructed the KG. Qing han built the BiLSTM-CRF model. Chang Liu and Jian-min Li finalized the experimental work, interpreted the results, and prepared the figures. Ying Chang provided much help in the KG's construction and the collation of results. Zhan Li and Chang Liu wrote the paper. Lingyong Cao and Shuyuan Lin provided medical guidance and participated in data analysis and interpretation.

Funding

This work was supported by the following: 1. Major project jointly built by the Department of Science and Technology of the State Administration of TCM and Zhejiang Provincial Administration of TCM (No. 2023014543) : Construction and multicenter application evaluation of an intelligent diagnosis and treatment system for T2DM based on graph neural network. 2. Special Project of TCM modernization in Zhejiang Province (No. 2022ZZ010): Construction of intelligent prescription recommendation model for type 2 diabetes mellitus based on knowledge; 3. Horizontal (enterprise-related) project of Zhejiang Chinese Medical University (No. 2020-ht-161): Construction and application of KG of the classical formulae; 4. Horizontal (enterprise-related) project of Zhejiang Chinese Medical University (No. 2020-ht-837): Research and application of knowledge reasoning of syndrome differentiation of the classical formulae based on KG.

Conflicts of Interest

Disclose any personal financial interests related to the subject matters discussed in the manuscript here. For example, authors who are owners or employees of Internet companies that market the services described in the manuscript will be disclosed here. If none, indicate with "none declared".

Abbreviations

KG: knowledge graph
TCM: traditional Chinese medicine
KGC: knowledge graph completion
KGE: knowledge graph embedding
RDF: resource description framework
CC: closeness centrality
MR: mean rank
MRR: mean reciprocal ranking

References

1. Wu,H, Liang Y, Li Q, et al. Analysis of the Characteristics of Dominant Diseases in Traditional Chinese Medicine: Based on 95 Diseases. Evidence-based complementary and alternative medicine : eCAM, 2022, 2022: 6972663.doi: 10.1155/2022/6972663. PMID: 35707474;

PMCID: PMC9192295.
2. Huang K, Zhang P, Zhang Z, et al. Traditional Chinese Medicine (TCM) in the treatment of COVID-19 and other viral infections: Efficacies and mechanisms: Efficacies and mechanisms. Pharmacology Therapeutics, 2021, (7647): 107843.
3. Liu GP, Wang YQ, Zhao NQ, et al. Study on the Diagnosis Agreement of Clinical Doctor of Traditional Chinese Medicine. World Science and Technology-Modernization of Traditional Chinese Medicine, 2010, 12(3): 358-62.
4. Lin SY, Zhu WP, Cao LY, et al. New Thought of Chinese Medicine Standardization Based on Characteristics of Classical Prescriptions Theory. Journal of Traditional Chinese Medicine, 2017, 58(24): 4.
5. Li LX, Yang F, Zhu ZX, et al Research Status and Development of Artificial Intelligence Syndrome Differentiation in Traditional Chinese Medicine. World Science and Technology-Modernization of Traditional Chinese Medicine, 2021, 23(11): 9.
6. Guo Q, Zhuang F, Qin C, et al. A Survey on Knowledge Graph-Based Recommender Systems. Scientia Sinica Informationis, 2020, 50(7): 937.
7. Tao YT, Chen YZ, Shao LY, et al. Construction and application of knowledge graph of traditional Chinese medicine. Beijing Journal of Traditional Chinese Medicine, 2022, 41(12): 6.
8. Huang HX, Wang XY, Gu ZW, et al. Research on Construction Technology and Development Status of Medical Knowledge Graph. Computer Engineering and Applications: 1-18.
9. Zhou XZ, Wu ZH, Yin AN, et al. Ontology development for unified traditinal Chinese medical language system. Artif Intell Med. 2004; 32:15–27.doi: 10.1016/j.artmed.2004.01.014. PMID: 15350621.
10. Long H, Zhu Y, Jia L, et al. An ontological framework for the formalization, organization and usage of TCM-Knowledge. BMC Med Inform Decis Mak. 2019; 19(Suppl 2):53.doi: 10.1186/s12911-019-0760-9. PMID: 30961578; PMCID: PMC6454592.
11. Guo MY, Zhou L, Sun Y. Application of the “Seven-Step Domain Ontology” Method in the Construction of a Traditional Chinese Medicine Diagnostic Reasoning Knowledge Base. World Science and Technology - Traditional Chinese Medicine Modernization, 2019, 21(12): 2646-2651.
12. Zhang YQ, Li ZY, Wang YH, et al. Construction of knowledge map on experience in TCM prescriptions of dermatology schools of Zhao bingnan and Zhu renkang. Chin J Libr Inf Sci Tradit ChinMed, 2021; 45(2): 1-5
13. Liu F, Wang MQ, Li LX, et al. Exploration on construction method of Knowledge Graph of veteran TCM physicians’ clinical experiences. China J Tradit Chin Med Pharm, 2021; 36(4): 2281-2285.
14. Zhong Y, Ru CL, Zang BL, et al. Research on Quality Control Methods of Traditional Chinese Medicine Preparation Process Based on Knowledge Graph. China Journal of Chinese Medicine, 2019, 44(24): 5269-5276.
15. Yu T, Li J, Yu Q, et al. Knowledge graph for TCM health preservation: Design, construction, and applications. Artif Intell Med. 2017 Mar;77:48-52.doi: 10.1016/j.artmed.2017.04.001. Epub 2017 Apr 21. PMID: 28545611.
16. Lu KZ. Construction and Application of Knowledge Graph based on Ancient Chinese Medical Literature, Beijing: Beijing Jiaotong University, 2020.
17. Yin D, Zhou L, Zhou YM, et al. Design and Research of "Graph Search Mode" for Traditional Chinese Medicine Formulae Knowledge Graph. Chinese Journal of Traditional Chinese Medicine Information, 2019, 26(08): 94-98.
18. Yu T, Jia LR, Liu J, et al. Research overview on traditional Chinese medicine language system. Chin J Libr Inf Sci Tradit ChinMed □ 2015 □ 39(6) : 56 □ 60 □
19. Wang M, Sun X, Liu J, et al. Visualization Analysis of Research Hotspots and Trends in

Traditional Chinese Medicine Xue Zhuo Theory based on CiteSpace Knowledge Graph. Chinese Journal of Medicine Guide, 2023, 20(12): 156-160. DOI: 10.20047/j.issn1673-7210.2023.12.35.
20.Wang XY, Yang T, Gao XY, Hu KF. Knowledge Graph Enhanced Transformers for Diagnosis Generation of Chinese Medicine. Chin J Integr Med. 2024 Mar;30(3):267-276.
21. Tao YT, Chen YZ, Shao LY, et al. Exploration on the Construction and Application of Traditional Chinese Medicine Knowledge Graph. Beijing Journal of Traditional Chinese Medicine, 2022, 41(12): 1387-1392.
22. Wang S, Li ZJ, Yang T, Hu KF, et al. Current Research Status and Development Trends of Traditional Chinese Medicine Knowledge Graph. Journal of Nanjing University of Traditional Chinese Medicine, 2022, 38(03): 272-278.
23. Sun MJ, Zhang D, Zheng MZ, et al. Traditional Chinese medicine aided diagnosis and treatment system for rheumatoid arthritis based on artificial intelligence. Pattern Recognit Artif In-tell, 2021, 34(4) : 343□352.
24. Fu ZX, Zhou P, Ren HY, et al. Inference Analysis of Integrative Diagnosis and Treatment for Acute Abdominal Pain Based on Knowledge Graph. Chinese Journal of Experimental Traditional Medical Formulae, 2023, 29(11): 190-199. DOI: 10.13422/j.cnki.syfjx.20230512.
25. Li P, Luo AJ, Min H. Establishment of prostate cancer diagnosis model based on big data of traditional Chinese medicine and graph convolutional network. Journal of Beijing University of Traditional Chinese Medicine, 2020, 43(12): 8.
26. Li DM, Qu JT, Tian ZW, et al. Knowledge-Based Recurrent Neural Network for TCM Cerebral Palsy Diagnosis. Evid Based Complement Alternat Med, 2022 Oct 12;2022:7708376. doi: 10.1155/2022/7708376. PMID: 36276852; PMCID: PMC9581687.
27.Dong WB, Sun SL, Yin MZ. Research and Development of Medical Knowledge Graph Reasoning. Journal of Frontiers of Computer Science & Technology, 2022, 16(06): 1193-213.
28.Bian H. Knowledge discovery and reasoning algorithm study in medical diagnose expert system. Yan shan University, 2016.
29.Bordes A, Usunier N, Garcia-duran A, et al. Translating Embeddings for Modeling Multi-relational Data; proceedings of the Neural Information Processing Systems; 2013 Dec 5-10; Lake Tahoe, United States. 2013.
30.Wang Z, Zhang JW, Feng JL, et al. Knowledge Graph Embedding by Translating on Hyperplanes. proceedings of the AAAI; 2014 July 27-31; Quebec City, Canada. 2014.
31.Lin YK, Liu Z, Sun MS, et al. Learning entity and relation embeddings for Knowledge Graph Completion; proceedings of the AAAI; 2015 Jan 25-30; Texas，United States. 2015.
32.Ji G, He S, Xu L, et al. Knowledge Graph Embedding via Dynamic Mapping Matrix; proceedings of the Meeting of the Association for Computational Linguistics & the International Joint Conference on Natural Language Processing; 2015 July 26-31; Beijing, China. 2015.
33.Sun Z, Deng Z H, Nie J Y, et al. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space; proceedings of the ICLR; 2019 Apr 18; New Orleans, United States. 2019.
34.Nickel M, Tresp V, Kriegel HP. A Three-Way Model for Collective Learning on Multi-Relational Data; proceedings of the International Conference on International Conference on Machine Learning; 2011, June 28-July 2; Washington, United States. 2011.
35.Yang BS, Yih WT, He XD, et al. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. ; proceedings of the ICLR; 2014 Apr 14-16; Banff, Canada. 2014.doi:10.48550/arXiv.1412.6575
36.Trouillon T, Welbl J, Riedel S, et al. Complex Embeddings for Simple Link Prediction. JMLRorg, 2016. doi: 10.48550/arXiv.1606.06357
37. Pujara J, Augustine E, Getoor L. Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short; proceedings of the Empirical Methods in Natural Language Processing, F, 2017 C. doi: 10.18653/V1/D17-1184

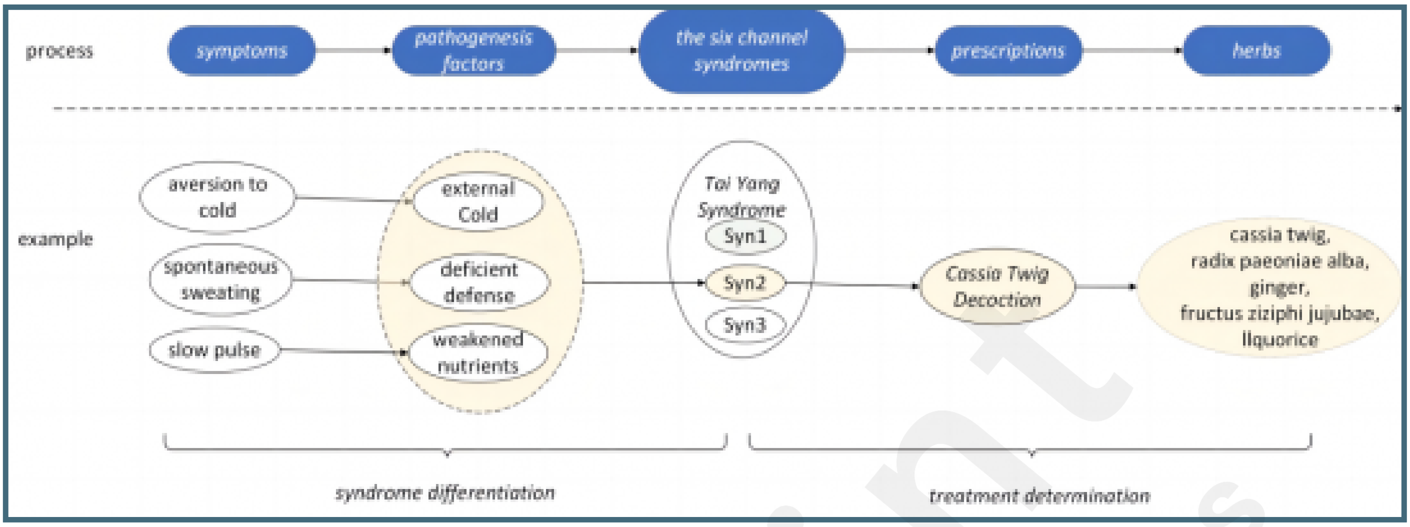
38. Xue BC, Zou L. Knowledge Graph Quality Management: A Comprehensive Survey. IEEE Transactions on Knowledge and Data Engineering. 2023; 35(5): 4969-4988. DOI: 10.1109/TKDE.2022.3150080.doi□10.1021/pr200824a
39. Xiu,X.,Qian,Q.,Wu,S. Construction of a Digestive System Tumor Knowledge Graph Based on Chinese Electronic Medical Records: Development and Usability Study. JMIR Medical Informatics, 2020, 8(10): e18287.doi□10.2196/18287
40. Lan Y, He S, Liu K, Zeng X, Liu S, Zhao J. Path-based knowledge reasoning with textual semantic information for medical knowledge graph completion. BMC Med Inform Decis Mak. 2021;21(Suppl 9):335.doi□10.1186/s12911-021-01622-7
41. Li X, Liu H, Zhao X, Zhang G, Xing C. Automatic approach for constructing a knowledge graph of knee osteoarthritis in Chinese. Health Inf Sci Syst. 2020;8(1):12.
42. Li L, Wang P, Yan J, et al. Real-world data medical knowledge graph: construction and applications. Artif Intell Med. 2020;103:101817.
43. Weng H, Chen J, Ou A, Lao Y. Leveraging Representation Learning for the Construction and Application of a Knowledge Graph for Traditional Chinese Medicine: Framework Development Study. JMIR Med Inform. 2022;10(9):e38414.doi: 10.2196/38414. PMID: 36053574; PMCID: PMC9482071.
44. Li L, Wang P, Wang Y, et al. A Method to Learn Embedding of a Probabilistic Medical Knowledge Graph: Algorithm Development. JMIR Med Inform. 2020;8(5):e17645.doi□10.2196/preprints.17645
45. Zheng S, Rao J, Song Y, et al. PharmKG: a dedicated knowledge graph benchmark for biomedical data mining. Brief Bioinform. 2021;22(4):bbaa344.doi: 10.1093/bib/bbaa344. PMID: 33341877.
46.Guo WF, Wu MH, Zhou ZY, et al. On ' pathogenesis syndrome factor '. Journal of Traditional Chinese Medicine, 2010, 51(05): 389-91.
47.Polanyi Michael. Personal Knowledge: Towards a Post-Critial Philosophy. London: Routledge,1958.
48. Lao N, Mitchell TM, Cohen WW. Random Walk Inference and Learning in A Large Scale Knowledge Base; proceedings of the Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL, F, 2011 C.
49. Lan Y, He S, Zeng X, et al. Path-based knowledge reasoning with textual semantic information for medical knowledge graph completion. 2021.doi: 10.1186/s12911-021-01622-7
50. Liu YQ. Researches and applications of knowledge graph and link prediction model for famous prescriptions of traditional Chinese medicine. Changchun: Northeast Normal University, 2021□
51. Shao, Xiang xiang; Hu, Kong fa; Dai, Caiyan. Knowledge Graph Reasoning of Famous Traditional Chinese Medicine for Lung Cancer Diagnosis and Treatment Based on RED-GNN. Journal of Software Guide, 2023, 22(03): 112-117.doi□10.1186/s12911-021-01622-7
52. Paulheim H. Identifying wrong links between datasets by multi-dimensional outlier detection J. eut edizioni università di trieste, 2014.DOI:Paulheim, Heiko (2014) Identifying Wrong Links between Datasets.
53. Chen ZK, Song X, Gao J, et al. Research Progress in Traditional Chinese Medicine Diagnosis and Treatment Based on Data Mining. Chinese Journal of Traditional Chinese Medicine. 2020;38(12):1-9. DOI:10.13193/j.issn.1673-7717.2020.12.001.
54. Xu HZ, Zhang T, Sun JL, et al. Application Progress of Association Rule Data Mining Methods in Traditional Chinese Medicine Research. Journal of Liaoning University of Traditional Chinese Medicine. 2013;15(12):131-134. DOI:10.13194/j.issn.1673-842x.2013.12.027.

55. Hu J. Analysis and Comparison of Several Typical Association Rule Algorithms. Modern Computer, 2011, (17): 15-7.
56. Laxminarayan P, Alvarez S A, Ruiz C, et al. Mining statistically significant associations for exploratory analysis of human sleep data. IEEE Transactions on Information Technology in Biomedicine A Publication of the IEEE Engineering in Medicine & Biology Society, 2006, 10(3): 440.doi：10.1109/TITB.2006.872065
57. Wang QG. Selection of Treatise on Febrile Diseases. Beijing: China Traditional Chinese Medicine Press, 2021.
58. Ding, LH, Sun, B, Shi, P. Empirical Research and Analysis of Complex Network Characteristics in Knowledge Graphs. Acta Physica Sinica, 2019, 68(12): 324-338.

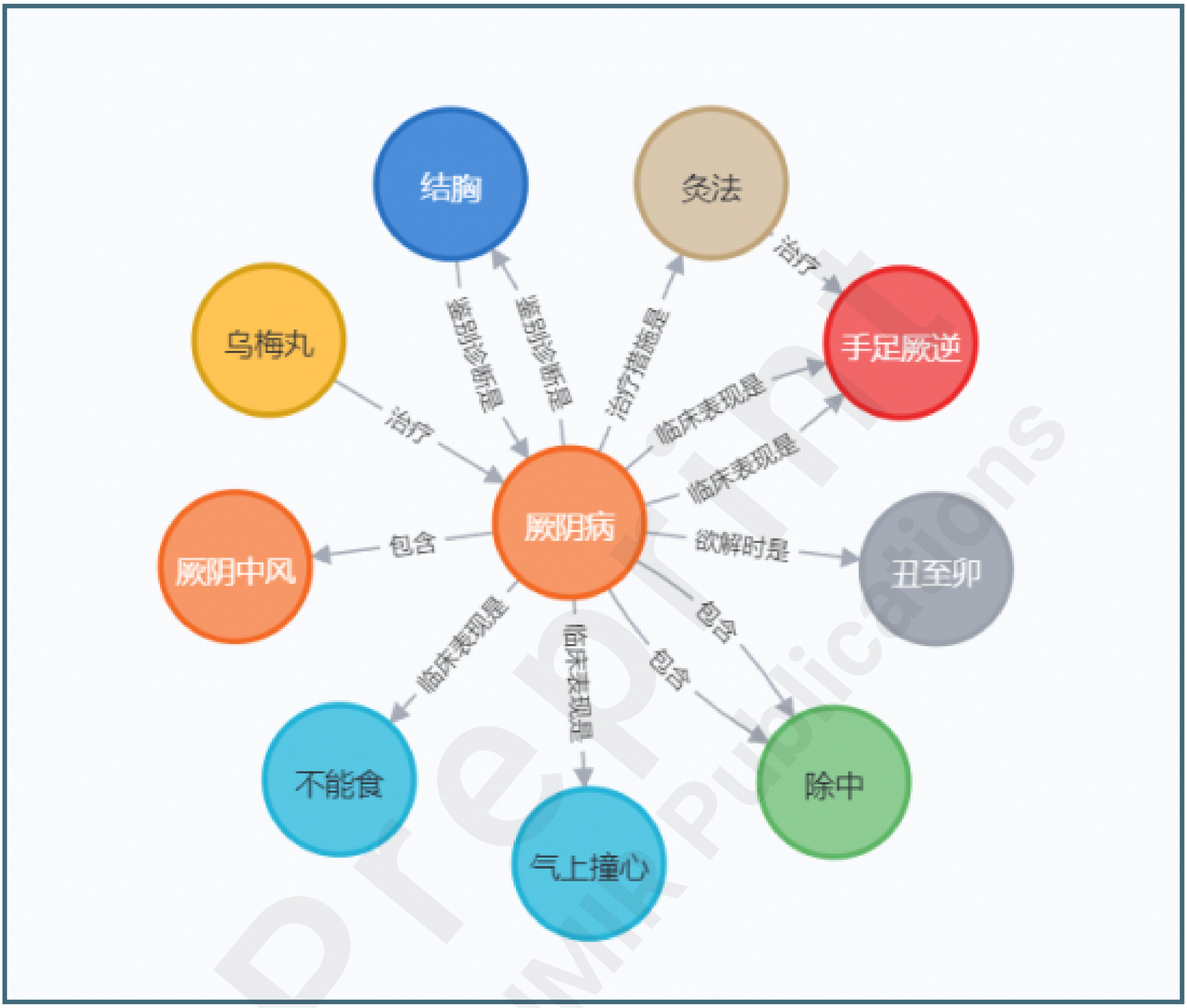
Supplementary Files

Figures

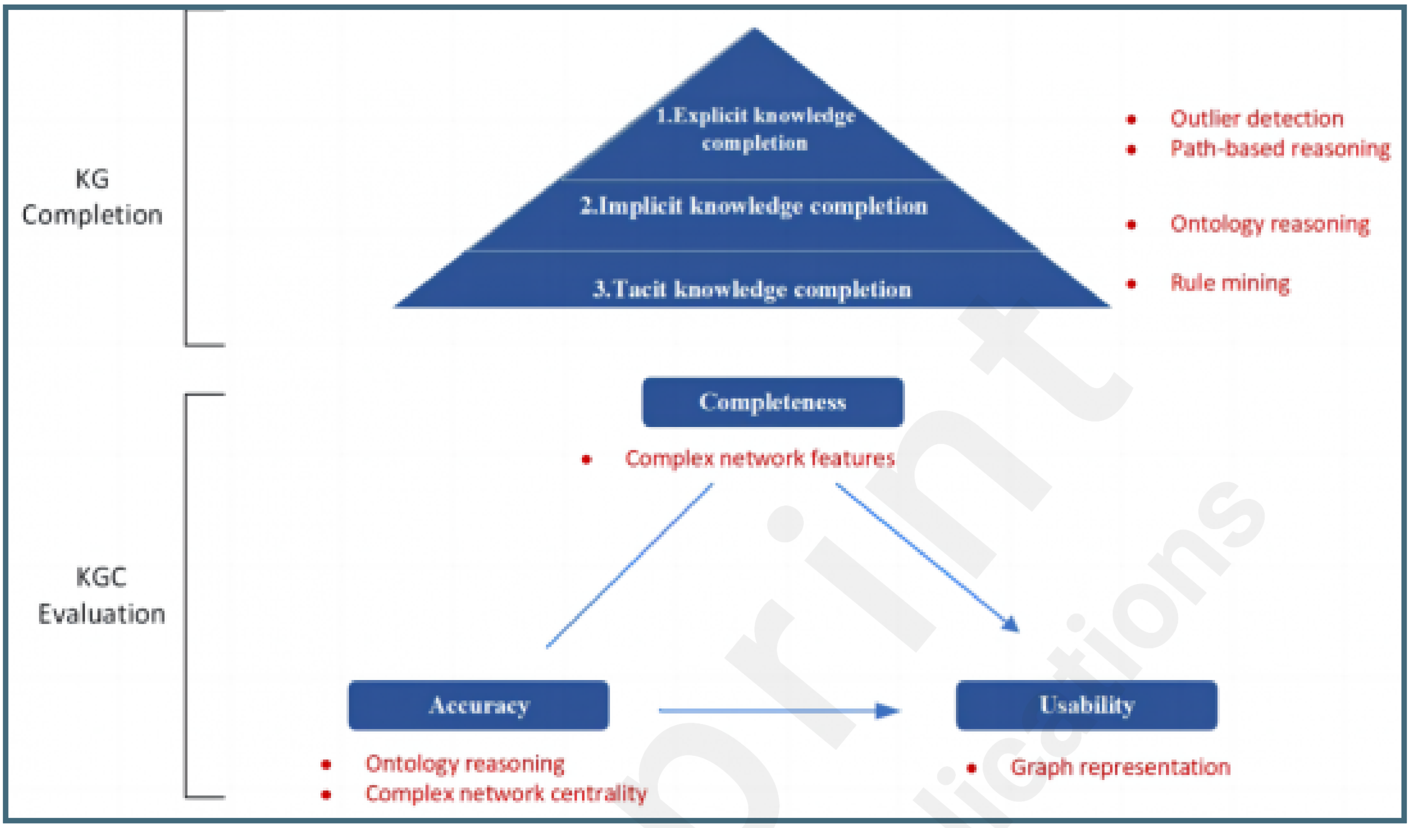
The Six Channel Syndrome Differentiation process.



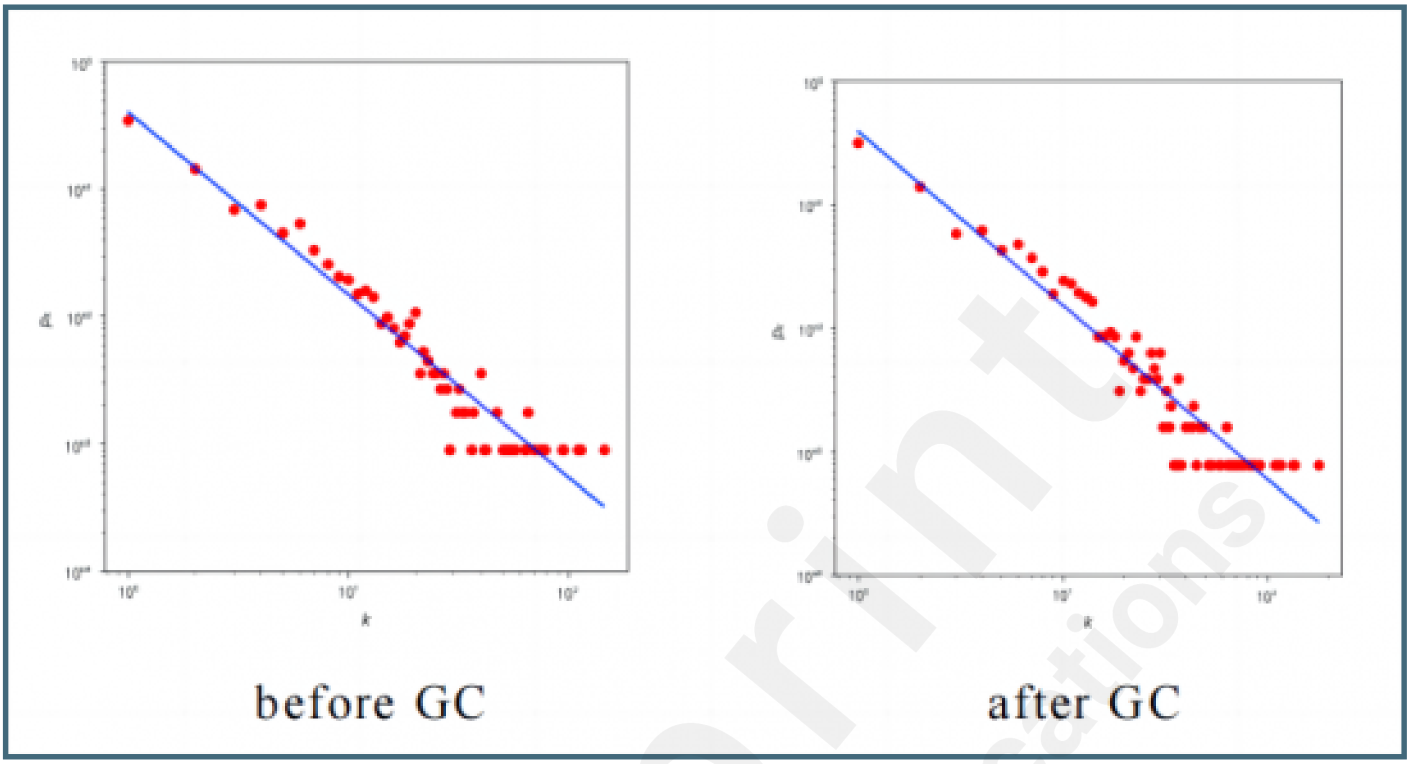
Part of KG based on Treatise on Febrile and Miscellaneous Diseases.



Methodology of research.



Distribution curve of the KG.


p_k
10^0
10^1
10^2
k
before GC
after GC