• Intelligent Ophthalmology •

Knowledge graph for traditional Chinese medicine diagnosis and treatment of diabetic retinopathy: design, construction, and applications

Li Xiao[1], Jing-Wei Wang[2], Cheng-Wu Wang[3], Ying Wang[3], Jun-Feng Yan[3], Qing-Hua Peng[4]

[1]School of Chinese Medicine, Hunan University of Chinese Medicine, Changsha 410208, Hunan Province, China
[2]Yiyang Central Hospital, Yiyang 413000, Hunan Province, China
[3]School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, Hunan Province, China
[4]Hunan Provincial Key Laboratory for Prevention and Treatment of Ophthalmology and Otolaryngology Diseases with Chinese Medicine, Hunan University of Chinese Medicine, Changsha 410208, Hunan Province, China
Correspondence to: Jun-Feng Yan and Qing-Hua Peng. Hunan University of Chinese Medicine, 300 Xueshi Road, Yuelu District, Changsha 410208, Hunan Province, China. junfengyan@hnucm.edu.cn; pqh410007@126.com
Received: 2024-06-14 Accepted: 2025-07-10

Abstract

● **AIM:** To develop a traditional Chinese medicine (TCM) knowledge graph (KG) for diabetic retinopathy (DR) diagnosis and treatment by integrating literature and medical records, thereby enhancing TCM knowledge accessibility and providing innovative approaches for TCM inheritance and DR management.

● **METHODS:** First, a KG framework was established with a schema-layer design. Second, high-quality literature and electronic medical records served as data sources. Named entity recognition was performed using the ALBERT-BiLSTM-CRF model, and semantic relationships were curated by domain experts. Third, knowledge fusion was mainly achieved through an alias library. Subsequently, the data layer was mapped to the schema layer to refine the KG, and knowledge was stored in Neo4j. Finally, exploratory work on intelligent question answering was conducted based on the constructed KG.

● **RESULTS:** In Neo4j, a KG for TCM diagnosis and treatment was constructed, incorporating 6 types of labels, 5 types of relationships, 5 types of attributes, 822 nodes, and 1,318 relationship instances. This systematic KG supports logical reasoning and intelligent question answering. The question answering model achieved a precision of 95%, a recall of 95%, and a weighted F1-score of 95%.

● **CONCLUSION:** This study proposes a semi-automatic knowledge-mapping scheme to balance integration efficiency and accuracy. Clinical data-driven entity and relationship construction enables digital dialectical reasoning. Exploratory applications show the KG's potential in intelligent question answering, providing new insights for TCM health management.

● **KEYWORDS:** diabetic retinopathy; traditional Chinese medicine; knowledge graph; intelligent question answering
DOI:10.18240/ijo.2025.11.01

Citation: Xiao L, Wang JW, Wang CW, Wang Y, Yan JF, Peng QH. Knowledge graph for traditional Chinese medicine diagnosis and treatment of diabetic retinopathy: design, construction, and applications. *Int J Ophthalmol* 2025;18(11):2011-2021

INTRODUCTION

Diabetic retinopathy (DR) is a common and serious microangiopathy resulting from the chronic effects of diabetes mellitus and remains a leading cause of preventable blindness[1]. With increasing DR trends worldwide, its prevention, treatment, and associated costs have become a significant concern for public health. The currently recommended western medicine (WM) treatments include anti-vascular endothelial growth factor therapies, retinal laser photocoagulation, and vitrectomy, most of which target advanced disease and have long-term side effects[2-4]. As a result, the management of DR remains challenging. Traditional Chinese medicine (TCM), a time-honored medical system in China, has been extensively utilized in the treatment of DR, yielding promising clinical outcomes[5-7]. Based on rich clinical practice and experimental research[8-10], the unique role of TCM in delaying DR progression and improving vision has attracted worldwide attention. Additionally, TCM offers the advantages of being cost-effective and having fewer side effects[11]. Therefore, harnessing the full potential of TCM in the prevention and treatment of DR holds significant promise.

Table 1 Related research on the construction of TCM knowledge graph

Study	Objective	Data source	KG construction	Strengths	Limitations	Application
Zhao *et al*[15]	Develop a TCM-KG for the diagnosis and treatment of DKD	Clinical guidelines, consensus, and real-world clinical data	Combine ontology and data mining	Association weights derived from data mining enriched DKD-KG relationships	Labor-intensive manual extraction	The discovery and sharing of diagnosis and treatment knowledge of DKD
Yin *et al*[16]	Build an intelligence question answering system for hepatitis B based on the hepatitis B KG	Xunyiwenyao website and medical records	A joint model leveraging a multi-head mechanism	Combined with the cutting-edge methods of current natural language processing technology	The amount of data is relatively small	The medical question answering system of hepatitis B disease
Cheng *et al*[17]	Construct a stroke-specific medical KG for general question answering applications	Vertical medical websites, crowdsourced encyclopedia websites, and the public knowledge base	Semi-automatic labeling method and dictionary construction	Similarity-based knowledge fusion and embedding-driven, iteratively optimized representations for precise entity associations	Limited types and quantity of data	The intelligent question answering system and auxiliary decision-making system of stroke

KG: Knowledge graph; TCM: Traditional Chinese medicine; DKD: Diabetic kidney disease.

However, the further development of TCM has encountered bottlenecks due to standardization issues. Therefore, to use TCM in a standardized and rational manner, it is crucial to synthesize the literature comprehensively and systematically. Nevertheless, the relevant literature is massive. Consequently, how to efficiently mine and utilize the literature has become a key issue. Modern information technology may offer a viable approach. The knowledge graph (KG), which transforms data into knowledge, is an emerging technology that can efficiently mine, represent, store, and present knowledge[12-13]. It is a large-scale relational knowledge base based on semantic networks and can capture and express complex relationships between domain concepts, making it suitable for the field of TCM. Therefore, this study adopted a semi-automatic approach to systematically organize DR-related literature, combined with clinical practice, to preliminarily verify the theory of TCM diagnosis and treatment for DR and construct the KG. Finally, we conducted research on question answering as a preliminary evaluation and application, which could provide a more reliable, convenient, and acceptable tool for healthcare workers and DR patients.

Since the introduction of the KG concept by Google in 2012, it has garnered significant attention and widespread application[14]. Among these, medicine is one of the most widely applied vertical fields for KGs. Recently, the application of KGs has garnered significant interest from the TCM community, prompting scholars to embark on research endeavors related to TCM knowledge graph. However, it is worth noting that the current research efforts are still nascent and encounter multiple challenges and constraints. To clarify current advancements and constraints, key studies[15-17] on TCM knowledge graphs are summarized in Table 1.

Overall, there are two main limitations to previous work on TCM knowledge graph construction. One focuses on the development of entity and relation extraction methods as well as model algorithms, which provides robust support for the discovery and reasoning of TCM-related diagnostic knowledge. However, due to the lack of standardized and comprehensive modeling strategies and processes, these studies have limited guidance for clinical applications[15,18]. On the other hand, there are studies on the construction of KGs tailored for specific domains, which pay more attention to data quality and often adopt manual methods. However, the efficiency and scalability of this approach are limited, making it difficult to handle large-scale and complex domain-specific KG construction. To address these issues, this study explores semi-automatic methods to improve the efficiency of KG construction. Simultaneously, a systematic modeling strategy and process are developed to construct a KG that integrates disease characteristics, top-level expertise, and clinical relevance. Through these efforts, we can anticipate that TCM knowledge graph will play a greater role in clinical practice, providing strong support for the modernization and internationalization of TCM.

MATERIALS AND METHODS

Ethical Approval The study protocol was approved by the hospital's ethics committee (approval number: HN-LL-YJSLW-2020-77).

Knowledge Graph Construction Framework KG construction includes schema layer and data layer. This study adopted a combination of "top-down" and "bottom-up" approaches to construct the KG (Figure 1). "Top-down" is a method for constructing the conceptual framework and relationships of the KG based on expert experience. It was used to construct the schema layer. "Bottom-up" was used to build the data layer, mainly through knowledge extraction, knowledge fusion, and the design of the KG's underlying storage model, to decompose concrete instance data and map them to the corresponding concept nodes, thereby complementing and improving the schema layer.

Design and Construction of Knowledge Graph

Schema layer Based on expert experience, a total of 6 types of concepts were ultimately defined, including disease, symptom, auxiliary examination, TCM syndrome, prescription, and drug. Five kinds of relationships between these concepts were formed, including contain, match, merge, diagnose, and treat. The specific design mode was shown in Figure 2.

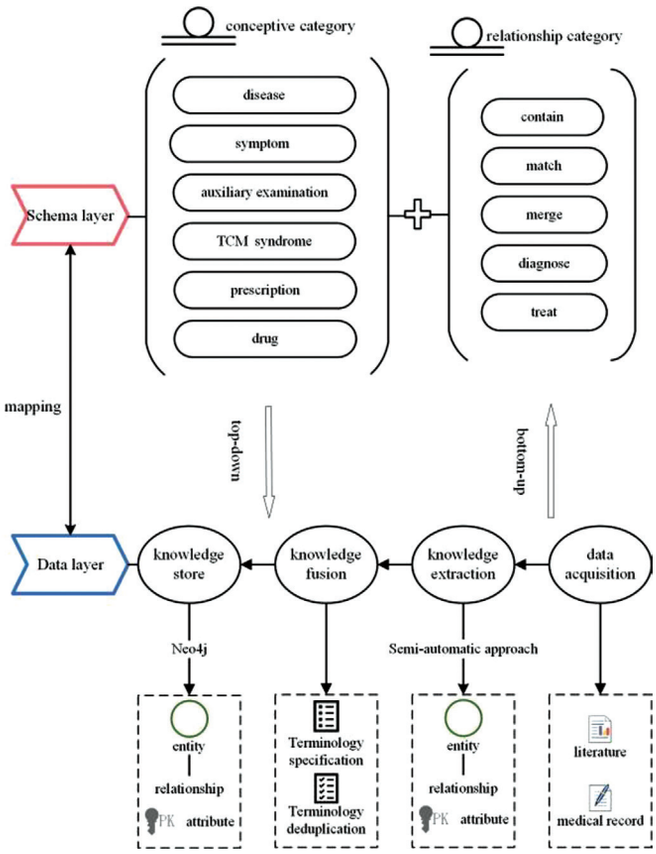

Figure 1 Workflow chart of the knowledge graph for TCM diagnosis and treatment of DR TCM: Traditional Chinese medicine; DR: Diabetic retinopathy.

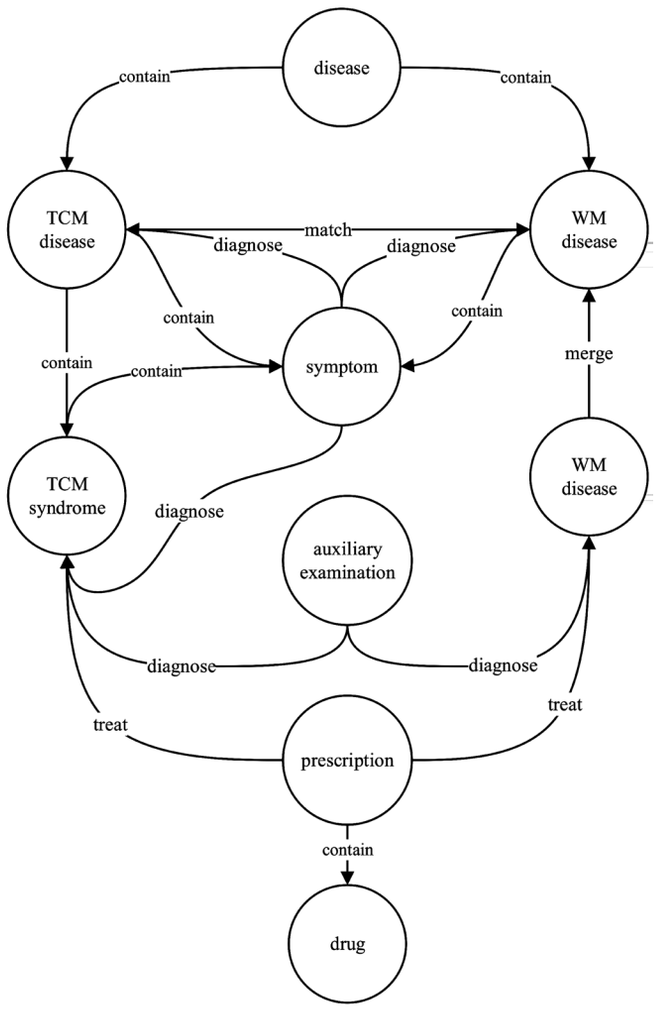

Figure 2 Schema-layer design diagram of the knowledge graph TCM: Traditional Chinese medicine; WM: Western medicine.

Data layer

1) Data sources

a) Medical record data In this study, we compiled a dataset comprising 6542 clinical records of DR patients in China. These records were collected from 2015 to 2021 and underwent a rigorous review process at the First Affiliated Hospital of Hunan University of Chinese Medicine. The data filtering criterion was that medical records lacking complete patient information, such as pulse condition, tongue condition, TCM diagnosis, and TCM prescription, were deleted. Finally, the manuscript used the medical records of 2904 patients that met the criteria. Each record contained the time of visit, age, sex, occupation, symptom description, medical history, anamnesis, diagnosis, TCM differentiation, prescription, and laboratory indicators, which are crucial for comprehensive analysis of DR patients.

b) Journal article The literature was retrieved from the China National Knowledge Infrastructure database and Wanfang database. The specific retrieval strategies were formulated based on the research objectives, including the publication type (journal article), topic (DR), time (unlimited), and language (Chinese). Subsequently, publication related to TCM was retrieved from the results through a full-text search (retrieved on September 11, 2021). We exported the title, author, organization, keywords, abstract, and other information in a custom format to an *.xls* file for further data processing and analysis. After de-duplication using Python, a total of 4614 articles were obtained.

c) TCM diagnosis and treatment plans, guidelines, and standards of DR It mainly included Chinese Multidisciplinary Expert Consensus on the Prevention and Treatment of Diabetic Eye Disease (2021 Edition)[19], Traditional Chinese Medicine Diagnosis and Treatment Standards for Diabetic Retinopathy[20], Diagnosis and Treatment Scheme for Diabetic Eye Disease (Diabetic Retinopathy)[21], Guidelines for the Prevention and Treatment of Diabetic Retinopathy in Traditional Chinese Medicine[22], Guidelines for Diagnosis and Treatment of Diabetic Retinopathy Combined with Disease and Syndrome(2021-09-24)[23]. These TCM diagnosis and treatment plans, guidelines, and standards for DR provide valuable references for defining concepts and relationships in the KG.

d) Book The relevant sections from *Ophthalmology of Traditional Chinese Medicine*[24] and *Fundus Diseases of Integrated Traditional and Western Medicine*[25], edited by

Prof. Peng QH, were incorporated into the study because they provide profound knowledge and valuable insights into TCM and integrated medicine approaches for treating DR.

2) Named entity recognition Entity extraction is one of the key tasks of knowledge extraction, which aims to obtain useful information from text. To ensure the quality and stability of knowledge, this study selected as many high-quality knowledge sources as possible, including journal literature, medical records, books, *etc*. However, these are unstructured texts. Based on our previous work[26], we compared the extraction performance of BiLSTM-CRF, BERT-BiLSTM-CRF and ALBERT-BiLSTM-CRF on ancient Chinese medicine books, and found that the ALBERT-BiLSTM-CRF model was more suitable for knowledge extraction of TCM descriptions.

ALBERT-BiLSTM-CRF model combines the strengths of A Lite Bidirectional Encoder Representations from Transformers (ALBERT), Bidirectional Long Short-term Memory (BiLSTM), and Conditional Random Field (CRF) for sequence labeling tasks in Natural Language Processing (NLP). ALBERT inherits the benefits of BERT, which is the ability to learn rich linguistic knowledge through unsupervised pre-training on large-scale corpora. However, with fewer parameters than BERT, ALBERT offers improved computational efficiency while maintaining high performance. Through fine-tuning, ALBERT can adapt to different downstream tasks, effectively extracting contextual representations of text. In addition, compared with currently dominant Word2Vec embedding method, ALBERT can better handle word polysemy[27]. BiLSTM is a type of recurrent neural network architecture, which is composed of two opposing LSTM networks. Compared with traditional recurrent neural network or unidirectional LSTM, BiLSTM can utilize contextual information more comprehensively[28]. CRF is a conditional random field that labels sequences by considering interactions between labels. It uses a global inference algorithm to optimize the labeling results of the entire sequence. The CRF model can capture contextual constraints and transitions between labels in sequence labeling tasks. The ALBERT-BiLSTM-CRF model combines the advantages of ALBERT, BiLSTM, and CRF, exhibiting significant strengths in NLP tasks, including efficient feature extraction capabilities, the ability to capture sentence structure and dependencies, global optimization of labeling results, and wide applicability[27]. Therefore, we used ALBERT-BiLSTM-CRF model for automatic recognition of named entities.

The included text data were randomly divided in a 7:3 ratio. 70% of the data were annotated using BIO tagging (Table 2). The specific rules of BIO tagging are as follows. Considering that prescriptions and drugs belong to the description of treatment, this study divided the text data into five different labels, including disease, symptom, auxiliary examination, TCM syndrome, and treatment. Meanwhile, non-entity tokens were labeled as "O".

Table 2 Named entity annotation methods

Label	Corresponding position
B-disease	Disease entity (beginning)
I-disease	Disease entity (not at the beginning)
B-symptom	Symptom entity (beginning)
I-symptom	Symptom entity (not at the beginning)
B-examination	Auxiliary examination entity (beginning)
I-examination	Auxiliary examination entity (not at the beginning)
B-TCMsyndrome	TCM syndrome entity (beginning)
I-TCMsyndrome	TCM syndrome entity (not at the beginning)
B-treatment	Treatment entity (beginning)
I-treatment	Treatment entity (not at the beginning)
O	Unnamed entity component

TCM: Traditional Chinese medicine.

The labeled data were randomly divided into the training set, the validation set, and the test set according to a 6:2:2 ratio, and the ALBERT-BiLSTM-CRF model was trained. The evaluation of the model's performance in the named entity recognition task employed standard evaluation metrics commonly used in multi-classification tasks: precision, recall, and the harmonic mean of precision and recall, known as the F1-score. To ensure the reliability of the experimental results, the model was repeatedly trained five times, and the average value was taken as the final evaluation result. The precision, recall, and F1-score of the model were 82%, 83%, and 83%, respectively (Figure 3). Finally, the trained model was used to predict the remaining 30% of the data, and the predictions were manually checked by trained personnel. The extracted data were stored in Excel. The results of entity extraction were cross-checked by two TCM doctors. In case of disagreement, a TCM ophthalmology professor adjudicated the results. In addition, entities were characterized by three core attributes: a Neo4j-generated ID assigned automatically, a normalized name derived from processed instances, and a category validated by domain experts.

3) Semantic relationship extraction and quantification Five types of relationships were defined based on expert experience: contain, match, merge, diagnose, and treat (Table 3). Furthermore, relationships were annotated with four core attributes: a Neo4j-generated ID assigned automatically, expert-defined name and type, and a clinically derived correlation (ranging from 0 to1) quantifying relationship strength. This structured framework enables both qualitative representation and quantitative analysis of entity relationships within the KG.

Knowledge fusion and alias resolution The data, sourced from multiple heterogeneous origins, necessitate knowledge fusion for seamless integration and association. Entity

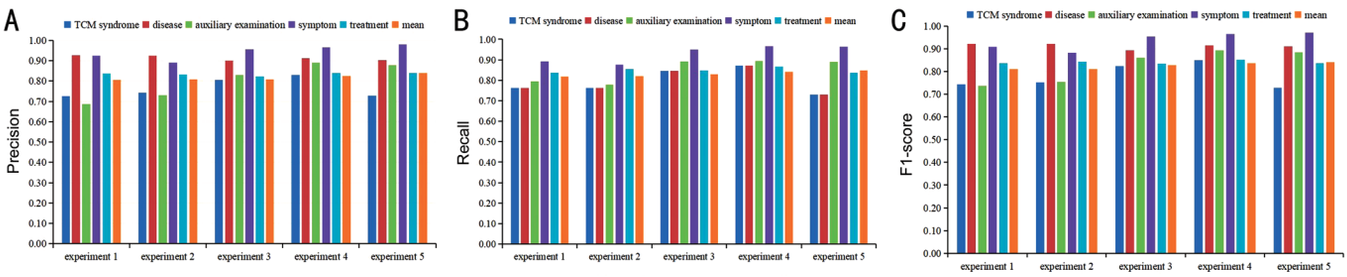

Figure 3 Experimental results of the ALBERT-BiLSTM-CRF model TCM: Traditional Chinese medicine.

Table 3 Definition of relationship types

Type	Description
Contain	An entity consists of several other entities
Match	One entity matches another entity, such as a WM disease name matching a TCM disease name
Merge	One entity coexists with another entity, such as a disease merging with a disease
Diagnose	One entity can support the diagnosis of another entity, for example, the results of an auxiliary examination can help doctors diagnose diseases
Treat	An entity has a therapeutic effect on another entity, such as a prescription treating a disease

TCM: Traditional Chinese medicine; WM: Western medicine.

Table 4 Representation of TCM diagnosis and treatment knowledge for DR in Neo4j

Element	Function	Object
Label	Describes conceptual entities in the schema layer	Concepts such as diseases, symptoms, TCM syndromes
Node	Describes specific entities	Concrete objects such as “diabetic retinopathy” and “blurred vision”
Relationship	Describes relationships between entities	Relationships such as “contain”, “treat”, “merge”, *etc.*
Attribute	Describes properties of entities and relationships	Properties such as “ID”, “name”, “category”, *etc.*

TCM: Traditional Chinese medicine; DR: Diabetic retinopathy.

alignment stands as the cornerstone of this fusion, ensuring clarity of meaning through entity disambiguation and coreference resolution. This study focused on standardizing relevant terminology to unite knowledge extracted from diverse sources. A comprehensive alias entity library was established, facilitating the fusion of multi-source data through entity mapping.

Knowledge representation and storage Compared with alternative knowledge representation methods, KG possesses superior representation capabilities, enabling a range of applications including semantic search, expert systems, intelligent question answering, decision support, personalized recommendations, *etc*. KG relies on a graph database as the fundamental storage engine, with Neo4j being a highly popular choice due to its stability, scalability, flexibility, portability, and impressive performance. Therefore, Neo4j was used in this study to store and represent TCM diagnosis and treatment knowledge for DR.

Labels, nodes, relationships, and attributes were used to represent knowledge in Neo4j (Table 4). In this study, the data were exported in CSV format for further processing. Using the py2neo module in Python, Cypher CREATE statements were executed to populate the Neo4j graph database. The data was stored in two main forms: entity-attribute-attribute value (with a one-to-one mapping) and entity-relationship-entity (with a one-to-many mapping).

RESULTS

After establishing the knowledge graph, a KG-based method for searching, browsing, and visualization was developed. This approach bridges gaps between knowledge silos and enhances the interconnectedness of knowledge resources in the TCM field. It helps users intuitively browse TCM knowledge at the concept level and discover potential connections between concepts, thereby better managing the complexity of the TCM health knowledge system. This study created 6 types of labels, 5 types of relationships, 5 types of attributes, 822 nodes, and 1318 relationship instances. The KG was also embedded into a WeChat official account platform, which provides authoritative, accurate, and comprehensive TCM health care knowledge for internet users. The applications of the KG are described as follows.

Knowledge Discovery TCM syndrome is a generalization of the pathological nature of the body at a certain stage in the course of disease development, encompassing etiology, disease location, disease nature, *etc*. Syndrome differentiation is pivotal for disease diagnosis, and treatment based on syndrome differentiation is the basic principle of TCM. Syndrome differentiation is a cognitive and practical process that establishes TCM syndromes by comprehending the disease dynamics. This entails gathering all pertinent disease information through the four diagnostic methods of

TCM, including symptoms. Subsequently, this information is analyzed and synthesized using TCM theories to discern the etiology, nature, location, and progression patterns of the disease, ultimately leading to the classification and diagnosis of a specific TCM syndrome. To illustrate, the following example was considered:

```
MATCH (n1:TCMsymptom {name: 'dryeye'})<-[r:contain]-(n2:TCMsyndrome)
RETURN n1, n2, r.weight
ORDER BY r.weight DESC
LIMIT 1
```

The results showed that a given TCM symptom may be related to multiple TCM syndromes, and the TCM syndrome with the strongest association (determined by the correlation) could be returned. For instance, if a patient presents with dry eye, the patient is most likely to be diagnosed with "liver and kidney deficiency, loss of nourishment in the eyes and collaterals syndrome" (Figure 4A). This finding is consistent with experts' clinical experience.

TCM treatment is individualized. Based on the theory of TCM syndrome differentiation and treatment, different TCM prescriptions are given for different syndrome types. A TCM prescription consists of a variety of Chinese medicinal herbs. If two prescriptions share some common Chinese medicinal herbs, it suggests that these common herbs may target similar pathological changes or related pathogenesis in these two syndrome types, which is worth exploring. In this study, specific Cypher queries can conduct a targeted and rapid search, taking the common herbs of Liu Wei Di Huang Pill and Jia Wei Shen Qi Pill as an example (Figure 4B), and the specific query was as follows:

```
MATCH (liuWei: Prescription {name: 'LiuWeiDiHuangPill'})-[r1]->(drug)<-[r2]-(jiaWei:Prescription {name: 'JiaWeiShenQiPill'})
RETURN liuWei, r1, drug, r2, jiaWei
```

Additionally, the frequent appearance of a Chinese medicinal herb in multiple representative prescriptions for treating DR suggests that this herb may possess a universal therapeutic effect in the treatment of DR or be valuable across different syndrome types. This insight paves the way for further exploration into the mechanisms of action and clinical applications of these medicines. Essentially, analyzing the frequency and occurrence of common Chinese medicinal herbs in representative prescriptions provides preliminary insights into their significance and function in DR management, serving as a reference point for subsequent comprehensive studies and clinical applications. In this study, specific Cypher queries facilitated the swift identification of high-frequency Chinese medicinal herbs (frequency threshold ≥5; Table 5). The query was:

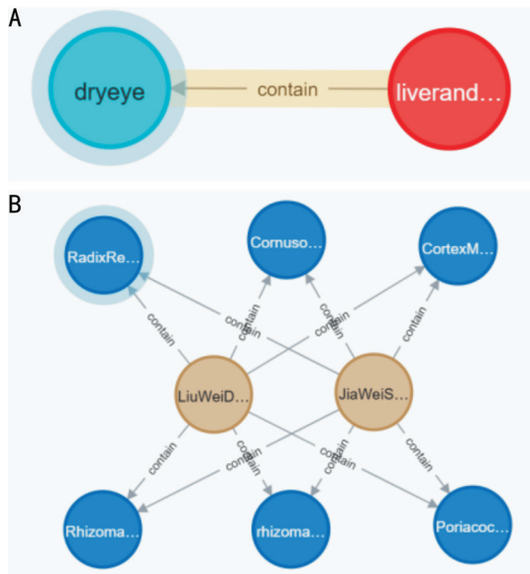

Figure 4 Knowledge graph of TCM diagnosis and treatment of DR from data mining A: TCM symptom–TCM syndrome; B: Drug prescription. Nodes represent entities; edges represent relationships; different colors represent different entities. Red nodes represent TCM syndrome types; green nodes represent TCM symptoms; blue nodes represent drugs; brown nodes represent prescriptions. TCM: Traditional Chinese medicine; DR: Diabetic retinopathy.

Table 5 High-frequency Chinese medicinal herbs based on relationship counts from data mining

Drug_name	Type of relationship	Number of relationships
Radix rehmanniae praeparata	Contain	10
Angelica sinensis	Contain	8
Liquorice	Contain	8
Poria cocos	Contain	8
Cornus officinalis	Contain	7
Rhizoma dioscoreae	Contain	7
Moutan radicis cortex	Contain	7
Ginseng	Contain	6
Schisandra chinensis	Contain	5
Rhizoma alismatis	Contain	5

```
MATCH (drug:Drug)<-[r]-(prescription:Prescription)
WITH drug, COUNT(r) AS relationship_count
WHERE relationship_count ≥5
RETURN drug.name AS drug_name, relationship_count
ORDER BY relationship_count DESC
```

Question Answering Based on the KG, we constructed an intelligent question answering system that provides reliable and convenient TCM diagnosis and treatment knowledge lookup services, as well as intelligent question answering services, for the majority of healthcare workers and patients. At the same time, the performance of the KG was also verified by a human–machine combined method guided by expert concerns.

Figure 5 Display of the intelligent question answering system interface TCM: Traditional Chinese medicine; DR: Diabetic retinopathy.

Table 6 Problem categories

Number	Problem category	Example
0	TCM syndromes of a certain disease	What are the TCM syndromes of DR
1	Symptoms of specific TCM syndromes	What are the main symptoms of "liver and kidney deficiency, loss of nourishment in the eyes and collaterals syndrome"
2	Diseases associated with a specific symptom	What may be the cause of blurred vision
3	Prescriptions for a specific TCM syndrome type	Which prescription can be used for "liver and kidney deficiency, loss of nourishment in the eyes and collaterals syndrome" of DR
4	Drug composition of a particular prescription	What is the composition of Ningxue Tang

TCM: Traditional Chinese medicine; DR: Diabetic retinopathy.

In this study, Jieba, a popular Chinese text segmentation tool[29], was used to preprocess the common questions. Jieba's robust segmentation capabilities allowed us to break down the text into meaningful units, facilitating the extraction of keywords. These keywords were then utilized to classify the questions into 5 categories (Table 6). Furthermore, ALBERT-BiLSTM-CRF model was used to identify entities, and then a Naive Bayes model was used to match question templates and generate the corresponding Cypher queries. Based on the identified entities, the query was executed on the KG of TCM diagnosis and treatment for DR to obtain the result. Finally, the answer was generated and returned to the user in a concise form, as shown in Figure 5. The answers were evaluated by two TCM experts and found to be consistent with clinical practice. After training, the macro-average and weighted-average accuracies reached 91% and 95% respectively. The macro average and weighted average of recalls were 93% and 95% respectively. The macro mean of F1 value was 91%, and the weighted mean was 95%. The test set contained 10,863 samples.

TCM Assisted Decision-making Digital syndrome differentiation of TCM is the core of intelligent assisted diagnosis of TCM[30]. In other words, TCM auxiliary decision-making aims to provide doctors with top N recommended TCM syndromes to facilitate follow-up treatment. Based on the correlation of "symptom–TCM syndrome", we can predict TCM syndromes according to patients' symptoms and rank them by likelihood. This approach enables doctors to make informed decisions by providing them with the most relevant TCM syndromes. For instance, when a DR patient presents with symptoms such as dizziness, tinnitus, dry eyes, chest tightness, pruritus, dry stool, a dull red tongue, and a thin pulse, the system can calculate the correlation degree for each TCM syndrome. Suppose the syndrome with the highest correlation degree is "liver and kidney deficiency with loss of nourishment in the eyes and collaterals", and this syndrome will be designated as the most recommended TCM syndrome type by the system.

DISCUSSION

Research Innovation and Significance This study integrates NLP technology with KG construction, enabling deeper mining and fusion of DR-related literature and real-world clinical data. The information was subsequently represented and stored in a machine-understandable format. This study proposes a

semi-automatic knowledge-mapping scheme. As a semantic network, the KG constructed in this study, empowered by DR diagnosis and treatment knowledge data, can solve a multitude of practical problems, laying a solid foundation for application fields such as intelligent question answering, intelligent query, and intelligent decision-making. From a medical resource perspective, the application exploration based on this constructed KG offers a novel approach for TCM health management. By enabling more efficient diagnosis and treatment strategies, it significantly contributes to the rational allocation and conservation of medical resources. In the realm of TCM education, the KG provides a structured and comprehensive knowledge system, facilitating the inheritance and dissemination of TCM theories and practical experiences. For clinical decision-making support, it offers evidence-based and intelligent recommendations, assisting medical professionals in formulating more accurate and personalized treatment plans for DR patients. Meanwhile, these applications also provide a preliminary quality evaluation of the KG.

In fact, question answering systems have been proposed since the advent of artificial intelligence and have been applied across various fields. However, their development in the medical domain–particularly systems that leverage real-world medical records as data sources–has been limited. The primary reason lies in the inherent complexity of medical knowledge and the lack of fixed patterns governing relationships between different entities[16]. KG serves as an effective means of knowledge organization, offering a solution to these challenges[31]. By harnessing these technologies, we can better utilize TCM, which has the potential to become a reliable complementary treatment for DR when combined with WM. In general, this study established a new knowledge base system for the TCM diagnosis and treatment KG for DR, along with a corresponding technical framework for knowledge organization, knowledge acquisition, and knowledge services.

Compared with other KGs, the KG constructed in this study has several unique characteristics. First, knowledge within the graph was derived from real-world medical records, which have been clinically verified, ensuring the practical applicability and reliability of the KG in a clinical setting. Second, the KG incorporated the unique logical system of TCM. By integrating these TCM-specific logical elements into the KG, it could better capture the essence of TCM knowledge and provide a more comprehensive understanding of DR from a TCM perspective. Additionally, this study integrated NLP with the KG for knowledge discovery and reasoning. NLP technology could efficiently process vast amounts of unstructured medical literature and records. In this study, the core technology of NLP, named entity recognition, was adopted to accurately extract key entities, including diseases, TCM syndromes, and other related terms. Building on our previous work, a cutting-edge hybrid model of ALBERT-BiLSTM-CRF was adopted. ALBERT can capture the domain-specific semantic features in medical texts, while the BiLSTM conducts in-depth modeling of context-dependent relationships, and the CRF layer performs global optimization of the label sequence. This combination significantly improves the recognition accuracy of complex terminologies. Experimental results also demonstrate that the model exhibits good performance on diverse TCM datasets, such as electronic medical records. This technology provides high-precision structured data support for constructing the TCM knowledge graph and for clinical decision-making assistance. Meanwhile, it validates the feasibility of integrating deep learning with knowledge in the TCM medical field, laying the foundation for subsequent multimodal medical information processing.

The entities extracted through named entity recognition and the relationships derived from expert experience were integrated into the KG, which serves as a structured knowledge repository. For knowledge discovery, the KG facilitates the exploration of hidden patterns and associations within medical knowledge. By traversing the graph and analyzing the relationships between different nodes, new knowledge can be discovered, such as potential treatment combinations for DR. In terms of reasoning, the KG can perform logical inferences based on the existing knowledge. For example, given a set of patient symptoms, the KG can use its stored knowledge and reasoning algorithms to infer the possible diagnosis and recommend appropriate treatment options. This process of structured knowledge discovery and reasoning is essential for intelligent medical applications, enhancing the efficiency and accuracy of medical decision-making.

Current Research Limitations Despite these significant achievements, the current study still has several limitations that require further improvement, primarily in terms of data constraints and model limitations, as detailed below.

Data limitations First, although a substantial amount of data was initially collected, the filtered dataset for model training was relatively small, which inherently restricts the sample size available for training. In addition, we tried to increase the training sample size, but the accuracy of TCM syndrome type identification decreased. This may be related to the uneven distribution of data, thus the data samples require further processing. The literature data employed in this study for KG construction covers a broad range of research topics, which are authoritative and representative. Nevertheless, the clinical data gathered lacked diversity, limiting its ability to encompass actual scenarios across various regions, medical standards, and patient demographics.

Model limitations The relationship extraction in this study

relied on expert experience. However, when dealing with massive unstructured data, manual extraction capabilities are limited, and subjectivity is inevitable. TCM data contains multiple types of intricate relationships, with many theoretical connections implicitly encoded with minimal explicit cues in the original text–a nuance comprehensible only to TCM experts. For instance, Traditional Chinese Medicine Language System lists and defines 96 basic semantic types and 58 semantic relations at the top level of ontology[32]. In fact, the number of types and the relationships among them are much greater than that. Currently, 1.27 million semantic relational links have been incorporated into Traditional Chinese Medicine Language System[32]. Despite advancements in relationship extraction techniques, ranging from simple rule-based methods to sophisticated hybrid parsers combining computational linguistics and machine learning, effective extraction of complex relationships among multitype entities in TCM remains challenging[33-35]. Extraction of implicit relations through collaboration between TCM experts and computer specialists may be a future trend.

In terms of model construction methods, existing KG completion techniques mainly include distance-based models[36], neural networks[37], and tensor decomposition methods[38]. Among these, translation-based models and tensor decomposition models are suitable for large-scale KG completion, whereas neural network models are suitable for KGs with complex relational structures. Traditional graph neural networks typically treat all neighboring nodes equally, ignoring hierarchical relational weights–a critical limitation in TCM knowledge graphs where relationships exhibit inherent asymmetry (*e.g.*, the nonequivalent roles of "sovereign, minister, assistant, and messenger" in herbal formulations). Moreover, existing knowledge completion methods do not make full use of entity semantic information. Although numerous approaches are being explored to address these issues[39-42], KG completion remains a significant challenge in the field. The primary challenges of current technologies include handling complex relationships, acquiring contextual semantics, capturing long-term dependencies between nodes, and improving model fusion and scalability. These areas represent critical directions for future research.

Certainly, TCM knowledge graph has emerged as a vibrant research focus, drawing increasing interest and investment from scholars worldwide. To fully capitalize on its potential within the realm of TCM, it is imperative to address and overcome a range of existing limitations and challenges. These encompass issues related to data quality and integration[18,43], the absence of standardized and normalized terminology[32], technical and methodological constraints[44], the absence of practical applications and validations[45], as well as ethical and privacy considerations[46].

Future Perspectives and Challenges Future research efforts should prioritize the effective integration of TCM data resources, establishing harmonized data standards and norms to enhance the quality and usability of the KG. During this process, it is necessary to carry out multicenter prospective clinical studies. By collaborating with multiple clinical institutions, more diverse and representative datasets can be collected. This will not only enrich the KG but also contribute to the formulation of more comprehensive data standards. In addition, with the advancement of deep learning technology, future research can improve the performance of NLP tasks in the field of TCM, including cross-lingual knowledge extraction, translation and representation. This will facilitate the internationalization of the TCM knowledge graph. Moreover, developing dynamic KG update mechanisms is vital to ensure that the latest research advancements and clinical practices are reflected in the KG.

As the availability of personal health data continues to grow, it is imperative to consider how to use these data for KG construction while safeguarding patient privacy. Through ongoing technological advancements, interdisciplinary collaboration, and the promotion of TCM globalization, there is a promising future for developing a more comprehensive, accurate, and practical TCM Knowledge graph. This enhanced KG has the potential to play a pivotal role in knowledge inheritance, education, clinical decision support, and new drug development, further advancing the field of TCM.

In summary, this study combined high-quality literature, guidelines, and real-world data to construct a KG for TCM diagnosis and treatment of DR, providing innovative perspectives for the representation and dissemination of TCM knowledge. The KG was built using a semi-automatic approach, which not only improved construction efficiency but also ensured knowledge accuracy. Furthermore, this study explored the application of the constructed KG by providing intelligent question answering services, offering a novel approach to TCM health management.

ACKNOWLEDGEMENTS

Authors' Contributions: Xiao L: Investigation, Conceptualization, Methodology, Software, Writing–original draft, Writing–review & editing; Wang JW: Methodology, Software, Writing–review & editing; Wang CW: Methodology; Wang Y: Investigation; Yan JF: Supervision, Writing–review & editing; Peng QH: Supervision, Writing–review & editing.

Foundations: Supported by Hunan Province Traditional Chinese Medicine Research Project (No.B2023043); Hunan Provincial Department of Education Scientific Research Project (No.22B0386); Research Project of Hunan Provincial Health Commission (No.20256982); Hunan University of

Traditional Chinese Medicine Campus Level Research Fund Project (No.2022XJZKC004).

Conflicts of Interest: Xiao L, None; **Wang JW,** None; **Wang CW,** None; **Wang Y,** None; **Yan JF,** None; **Peng QH,** None.

REFERENCES

1 Simó R, Simó-Servat O, Bogdanov P, *et al*. Diabetic retinopathy: role of neurodegeneration and therapeutic perspectives. *Asia Pac J Ophthalmol* (*Phila*) 2022;11(2):160-167.

2 Zhao N, Guan J, Cai N, *et al*. Efficacy of intravitreal conbercept combined with panretinal photocoagulation for severe nonproliferative diabetic retinopathy without macular edema. *Int J Ophthalmol* 2022;15(4):615-619.

3 Lee WA, Shao SC, Liao TC, *et al*. Effect modification by indication to the risks of major thromboembolic adverse events in patients receiving intravitreal anti-vascular endothelial growth factor treatment: a population-based retrospective cohort study. *BioDrugs* 2022;36(2):205-216.

4 Kupis M, Samelska K, Szaflik J, *et al*. Novel therapies for diabetic retinopathy. *Cent Eur J Immunol* 2022;47(1):102-108.

5 Huai B, Huai B, Su Z, *et al*. Systematic evaluation of combined herbal adjuvant therapy for proliferative diabetic retinopathy. *Front Endocrinol* 2023;14:1157189.

6 Xiao L, Yang YJ, Liu Q, *et al*. Visualizing the intellectual structure and recent research trends of diabetic retinopathy. *Int J Ophthalmol* 2021;14(8):1248-1259.

7 Du J, Mao Y, Xu Y, *et al*. Shuangdan Mingmu capsule for diabetic retinopathy: a systematic review and meta-analysis of randomized controlled trials. *Evid Based Complement Alternat Med* 2023;2023:4655109.

8 Zhao ZH, Xu M, Fu C, *et al*. A mechanistic exploratory study on the therapeutic efficacy of astragaloside IV against diabetic retinopathy revealed by network pharmacology. *Front Pharmacol* 2022;13:903485.

9 Fang Y, Shi K, Lu H, *et al*. Mingmu Xiaomeng Tablets restore autophagy and alleviate diabetic retinopathy by inhibiting PI3K/Akt/mTOR signaling. *Front Pharmacol* 2021;12:632040.

10 Du A, Xie Y, Ouyang H, *et al*. Si-Miao-yong-an decoction for diabetic retinopathy: a combined network pharmacological and *in vivo* approach. *Front Pharmacol* 2021;12:763163.

11 Li X, Zhang J, He R, *et al*. Effect of Chinese herbal compounds on ocular fundus signs and vision in conventional treated-persons with non-proliferative diabetic retinopathy: a systematic review and meta-analysis. *Front Endocrinol (Lausanne)* 2022;13:977971.

12 Gao Z, Ding P, Xu R. KG-Predict: a knowledge graph computational framework for drug repurposing. *J Biomed Inform* 2022;132:104133.

13 Lu M, Zhang Y, Zhang S, *et al*. Knowledge-aware patient representation learning for multiple disease subtypes. *J Biomed Inform* 2023;138:104292.

14 Singhal A. Introducing the knowledge graph: things, not strings. Official Google Blog; 2012.

15 Zhao X, Wang Y, Li P, *et al*. The construction of a TCM knowledge graph and application of potential knowledge discovery in diabetic kidney disease by integrating diagnosis and treatment guidelines and real-world clinical data. *Front Pharmacol* 2023;14:1147677.

16 Yin Y, Zhang L, Wang Y, *et al*. Question answering system based on knowledge graph in traditional Chinese medicine diagnosis and treatment of viral hepatitis B. *Biomed Res Int* 2022;2022:7139904.

17 Cheng B, Zhang J, Liu H, *et al*. Research on medical knowledge graph for stroke. *J Healthc Eng* 2021;2021:5531327.

18 Xiao W, Zhang M, Zhao D, *et al*. TCMKD: From ancient wisdom to modern insights-a comprehensive platform for traditional Chinese medicine knowledge discovery. *J Pharm Anal* 2025;15(6):101297.

19 Diabetic Retinopathy Group of Chinese Diabetes Society. Chinese multidisciplinary expert consensus on the prevention and treatment of diabetic eye disease (2021 edition). *Chin J Diabetes Mellitus* 2021;13(11):1026-1042.

20 Duan JG, Jin M, Jie CH, *et al*. Traditional Chinese medicine diagnosis and treatment standards for diabetic retinopathy. *World Journal of Integrated Traditional and Western Medicine* 2011;6(7):632-637.

21 National Administration of Traditional Chinese Medicine "Eleventh Five Year Plan Experts" Cooperation Group. *Diagnosis and treatment scheme for diabetic eye disease (diabetic retinopathy)*. Beijing:China Press of Traditional Chinese Medicine, 2013.

22 Duan JG, Jin M, Jie CH. Guidelines for the prevention and treatment of diabetic retinopathy in traditional chinese medicine. *Chinese Medicine Modern Distance Education of China* 2011;9(4):154-155.

23 Chen Q, Ni Q, Liu Y. Guidelines for diagnosis and treatment of diabetic retinopathy combined with disease and syndrome (2021-09-24). *World Chinese Medicine* 2021;16(22):3270-3277.

24 Peng QH. *Ophthalmology of Traditional Chinese Medicine*. Beijing:*China Press of Traditional Chinese Medicine*, 2021.

25 Peng QH. *Fundus Diseases of Integrated Traditional and Western Medicine*. Beijing:People's Military Medical Press, 2011.

26 Wang JW, Xiao L, Luo JW, *et al*. Research on Named Entity Recognition Based on Treatise on Febrile Diseases. *Computer & Digital Engineering* 2021;49(8):1584-1587.

27 Wei J, Hu T, Dai J, *et al*. Research on named entity recognition of adverse drug reactions based on NLP and deep learning. *Front Pharmacol* 2023;14:1121796.

28 Shahid F, Zameer A, Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos Solitons Fractals* 2020;140:110212.

29 Yang BX, Chen P, Li XY, *et al*. Characteristics of high suicide risk messages from users of a social network-sina weibo "tree hole". *Front Psychiatry* 2022;13:789504.

30 Wei CF, Yan JF. Research on construction of intelligent acquisition and analysis system of supporting medical cases for traditional Chinese medicine digital syndrome differentiation. *Journal of Hunan University of Chinese Medicine* 2020;40(1):70-74.

31 Freidel S, Schwarz E. Knowledge graphs in psychiatric research: Potential applications and future perspectives. *Acta Psychiatr Scand* 2025;151(3):180-191.

32 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.

33 Zhou D, Zhong D, He Y. Biomedical relation extraction: from binary to complex. *Comput Math Meth Med* 2014;2014:298473.

34 Zeng Z, Tong L, Li B, *et al*. TCMSF: a construction framework of traditional Chinese medicine syndrome ancient book knowledge graph. *Methods Inf Med* 2024;63(5-06):183-194.

35 Zhang T, Huang Z, Wang Y, *et al*. Information extraction from the text data on traditional Chinese medicine: a review on tasks, challenges, and methods from 2010 to 2021. *Evid Based Complement Alternat Med* 2022;2022:1679589.

36 Zhao Y, Bollegala D, Hirose S, *et al*. Community knowledge graph abstraction for enhanced link prediction: a study on PubMed knowledge graph. *J Biomed Inform* 2024;158:104725.

37 Zhao X, Chen L, Chen H. A weighted heterogeneous graph-based dialog system. *IEEE Trans Neural Netw Learn Syst* 2023;34(8): 5212-5217.

38 Jia T, Yang Y, Lu X, *et al*. Link prediction based on tensor decomposition for the knowledge graph of COVID-19 antiviral drug. *Data Intell* 2022;4(1):134-148.

39 Biswas S, Mitra P, Rao KS. Relation prediction of co-morbid diseases using knowledge graph completion. *IEEE/ACM Trans Comput Biol Bioinform* 2021;18(2):708-717.

40 Dai G, Wang X, Zou X, *et al*. MRGAT: Multi-Relational Graph Attention Network for knowledge graph completion. *Neural Netw* 2022;154:234-245.

41 Zhang M, Geng G, Zeng S, *et al*. Knowledge graph completion for the Chinese text of cultural relics based on bidirectional encoder representations from transformers with entity-type information. *Entropy* (*Basel*) 2020;22(10):1168.

42 Lan Y, He S, Liu K, *et al*. Path-based knowledge reasoning with textual semantic information for medical knowledge graph completion. *BMC Med Inform Decis Mak* 2021;21(Suppl 9):335.

43 Shataer D, Cao S, Liu X, *et al*. Application of large language models in traditional Chinese medicine: a state-of-the-art review. *Am J Chin Med* 2025;53(4):973-997.

44 Guo ZH, Liu QP, Zou BJ. Research on knowledge reasoning of TCM based on knowledge graphs. *Digital Chinese Medicine* 2022;5(4):386-393.

45 Zhang Z, Fang M, Wu R, *et al*. Large-scale biomedical relation extraction across diverse relation types: model development and usability study on COVID-19. *J Med Internet Res* 2023;25:e48115.

46 Wiedemann P. Artificial intelligence in ophthalmology. *Int J Ophthalmol* 2023;16(9):1357-1360.