Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TCMCPD: An Integrated TCM Case and Pharmacology Database

Junbai Chen
Peking University
Ye Ma
Beijing University of Chinese Medicine
BuTian Zhao
Beijing University of Chinese Medicine
RuoJia Wang
Beijing University of Chinese Medicine
Guo Chen
Beijing University of Chinese Medicine
FengYing Guo
guofy@bucm.edu.cn
Beijing University of Chinese Medicine

Research Article

Keywords: Traditional Chinese medicine case, Named Entity Recognition, Traditional Chinese medicine database, Knowledge Graph, Natural Language Processing

Posted Date: May 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9420486/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Traditional Chinese medicine (TCM) cases, serving as a crucial vehicle for TCM diagnostic and therapeutic experience, document the clinical reasoning, logic of syndrome differentiation, and treatment protocols of renowned senior TCM practitioners. However, classical medical cases are predominantly documented as unstructured text with severe information fragmentation, making systematic integration and knowledge mining challenging. At present, research on the construction of medical case databases remains relatively scarce. Existing databases primarily focus on case collection for single or rare diseases, and are predominantly composed of clinical disease data from Western medicine, lacking multidisciplinary clinical data in TCM; Concurrently, research related to network pharmacology databases is also solely focused on the research and analysis of TCM, drugs, targets, and diseases.

Methods

Taking the 6 volumes of *Essence of Medical Cases of Modern Famous TCM Physicians in China* as the data source, we performed digitization processing on the contents, integrated deep learning algorithms to conduct structured processing of entities including patient medical records, TCM syndromes,TCM symptoms, prescriptions and TCM herbs, and achieved standardization by mapping and aligning with the self-constructed TCM knowledge base. Meanwhile, data on ingredients, targets and genes were obtained from authoritative publicly available domestic and international databases. The TCMCPD was constructed based on MySQL, with a complete association relationship network established, and a Web-based interactive visualization platform was developed.

Results

We constructed a database integrating clinical diagnosis and treatment and pharmacology, with a complete relational chain of “Patient-Visit-TCM symptom-TCM syndrome-TCM herb-ingredient-target”. The TCMCPD contains 5172 patients, 12900 medical records, 1717 symptoms, 233 TCM syndromes, 499 types of TCM herbs, 19595 ingredients, and 5235 targets, with all data visualized through a knowledge graph.

Conclusions

TCMCPD not only integrates the basic information and diagnosis and medica records of TCM cases, but also incorporates data including chemical ingredients and targets of Chinese materia medica from existing TCM databases. It reveals the internal logic and diagnosis–treatment patterns of TCM cases, provides data support for clinical practice, scientific research analysis and research and development of

Chinese materia medica, empowers the application of TCM artificial intelligence (AI), and facilitates the integration of traditional Chinese medicine and modern medicine.

Background

Traditional Chinese medicine (TCM) cases are important vehicles for reflecting the thoughts of renowned veteran TCM physicians. Using technologies like big data and Artificial Intelligence (AI) to study these medical records supports the inheritance and promotion of TCM culture, enriches TCM theory, guides clinical practice, and advances TCM informatization and internationalization[1, 2]. However, syndrome differentiation and treatment still rely heavily on physician experience, mainly due to limited integration of structured medical records with pharmacological data and weak links between TCM diagnosis and treatment and modern pharmacological mechanisms. Medication decisions remain somewhat subjective[3, 4], and managing large volumes of unstructured medical records and clinical data in a standardized way has become an urgent challenge in TCM modernization[5, 6].

Existing studies have shown that medical case database models are becoming increasingly diverse, covering TCM inheritance, single-case management, rare disease surveillance, clinical reasoning. Firstly, TCM inheritance databases focus on preserving the clinical experience of renowned veteran TCM physicians, which can support TCM clinical research and decision-making by integrating syndrome differentiation and treatment with prescribing rules[7]. For instance, the “TCM Medical Case Intelligence Platform”[8]and the “Ancient and Modern Medical Case Cloud Platform”[9]compile clinically validated classic TCM cases, document the diagnostic thinking and prescribing patterns of the physicians and establish a reliable TCM clinical knowledge base. Secondly, Single-case databases address the fragmentation of clinical evidence. As an example, Iwakabe et al.[10] constructed an electronic clinical database by aggregating published single-case studies, improving reuse of clinical evidence and allowing physicians to rapidly retrieve intervention processes and treatment outcomes from similar cases, thus supporting inheritance of clinical experience. Thirdly, rare disease databases emphasize multi-source data integration and dynamic epidemiological surveillance. For example, Zhang et al.[11, 12] combined query data from large online search engines with confirmed case data from the national reporting system, achieving dynamic surveillance of rare diseases and helped predict the potential patient population by using search frequency to develop models. Finally, clinical case databases are designed for clinical reasoning use artificial intelligence to analyze large volumes of unstructured case texts. For example, Zack et al.[13] applied natural language processing (NLP), unsupervised machine learning, and ontology mapping to extract medical terms from case texts and cluster them by topic, aiming to reveal the implicit relationships among symptoms, pathophysiology, and diagnosis, whose research can support advanced clinical reasoning training, evidence-based medical education, and complex clinical decision-making.

Meanwhile, based on Li et al.’s[14] concept of TCM network pharmacology derived from systems biology, research on TCM network pharmacology databases has developed rapidly. Existing representative databases fall into two major types: integrated platforms for network analysis of TCM formulas that also

incorporate toxicological and adverse-effect information, such as TCM-Mesh[15]; and knowledge bases designed to mine “ingredient–target–disease” relationships, such as TCMSP[16], HIT[17], TCM Database@Taiwan[18], and HERB[19]. These databases not only provide robust data for drug innovation and efficacy evaluation, but are also widely used for active compound screening, potential target prediction, analysis of complex disease mechanisms, investigation of compatibility patterns in TCM formulas, network toxicology assessment, and the construction and systematic analysis of visualized networks[20].

Although the development of these databases has promoted the modernization and informatization of TCM, several limitations remain. First, existing medical case databases rely on relatively limited data sources, most of which are derived from Western medical records, whereas the systematic integration of TCM clinical data remains insufficient, particularly the collection, processing, and mining of case records from ancient TCM texts. Second, there are relatively few TCM medical case inheritance databases, most of which focus on the case texts, without incorporating related data on pharmacology, herbal ingredients, and mechanisms of action. Third, current network pharmacology databases are mainly oriented toward pharmaceutical research and lack patient diagnostic and treatment information, making it difficult to fully trace the relationships among symptoms, syndrome differentiation, herbal prescriptions, and therapeutic targets.

Therefore, our study used medical records of modern renowned TCM physicians as the primary data source and integrated them with data obtained from public databases to construct a TCM Case & Pharmacology Database (TCMCPD) covering “Patient–TCM symptom–TCM syndrome–TCM herb–Ingredient–Target” relationships. Based on it, knowledge graphs were employed for visual presentation, revealing the complete link from clinical to molecular mechanisms. The resulting database provides physicians with objective evidence to support syndrome differentiation and treatment, offers researchers high-quality data for investigating the mechanisms and compatibility patterns of TCM herbs, supplies a structured knowledge base for AI applications, and serves as an educational resource for the inheritance of TCM knowledge.

Methods

Based on the patient symptoms, syndromes and medication information documented in the structured data of ancient TCM cases, combined with the collection of information on TCM symptoms, TCM syndromes, as well as ingredients and targets of TCM herbs from existing public databases, we constructed a TCM case database covering both diagnosis and pharmacology. The architecture diagram of the database construction is shown in Fig. 1.

Data Collection and Processing

TCM Cases Data Collection and Processing

The database we constructed consists of seven core modules: patient, visit, TCM syndrome, TCM symptom, TCM herb, ingredient, and target. Patient information, Medica records (including the original text of TCM cases, tongue manifestations, pulse conditions, and therapeutic methods), TCM syndromes, symptom descriptions, and names of TCM herbs used were sourced from the 6 volumes of *Essence of Medical Cases of Modern Famous TCM Physicians in China*. A total of 5172 TCM cases from 434 renowned veteran TCM physicians were included, covering specialties such as TCM internal medicine, TCM surgery, and TCM gynecology (Fig. 1). We developed an Optical Character Recognition (OCR) program using Python 3.7 to convert the extracted PDF text of TCM cases into TXT text format, and performed manual verification and correction against the original text. Meanwhile, a unique PatientID was assigned to each TCM case. The text of multiple consultation records for the same patient was segmented chronologically, and a unique VisitID was assigned to each consultation record text. Finally, the verified and corrected TCM case data were stored item by item in Excel files according to each individual consultation visit.

To improve the preliminary screening efficiency and structured processing capability of unstructured text data, we developed the Named Entity Recognition (NER) model TCMYianBERT-BDM-CRF [21]in our previous research. This model implements structured processing of the above TCM case data by integrating three neural network models: Bi-directional Long Short-Term Memory (BiLSTM)[22], Gated-Dilated Convolutional Neural Networks (DGCNN)[23], and Multi-Head Attention (MHA)[24]. Experimental comparisons reveal that this model exhibits excellent performance on the TCM case dataset, with a Precision of 71.47%, a Recall of 76.00%, and an F1-score of 73.67%, significantly outperforming baseline models such as the traditional BERT-BiLSTM-CRF[25].

To ensure the reliability and accuracy of the data, this study first utilized the TCMYianBERT-BDM-CRF model to conduct automated entity recognition and information extraction from 12,900 visit records, extracting key entities including TCM symptoms, syndromes and TCM herbs. On this basis, experts in the field of TCM were organized to conduct comprehensive manual review and correction on the model extraction results, so as to ensure the accuracy and semantic consistency of entity recognition and relation extraction. Meanwhile, standardized mapping and terminology alignment were performed on the identified entities in accordance with current industry standards and authoritative TCM monographs, including *Classification and Codes of Traditional Chinese Medicine Diseases and Syndromes* (GB/T 15657 – 2021), *Classification and Codes of TCM Tongue Manifestation Diagnosis Information* (T/CIATCM 010-2019), *Classification and Codes of TCM Pulse Manifestation Diagnosis Information* (T/CIATCM 011-2019), *Classification and Codes of Basic Clinical Symptom Information of Traditional Chinese Medicine* (T/CIATCM 020-2019), *Science of TCM Diagnosis*, *Chinese Materia Medica*, the series of *Terms of Traditional Chinese Medicine* issued by the China National Committee for Terms in Sciences and Technologies, *Science of Differential Diagnosis of TCM Symptoms*, *Studies on TCM Symptomatology*, and *Clinical Terminology of TCM Diagnosis and Treatment*, so as to achieve standard unification and semantic standardization at the datalevel.

Public Data Collection

Detailed information on TCM syndromes, TCM symptoms,TCM herbs, and data on ingredients and targets were integrated from the SymMap database[26].It covers basic Chinese materia medica information from the *Pharmacopoeia of the People's Republic of China*, chemical constituent data from the TCMID and TCMSP databases, constituent targets of Chinese materia medica from the HIT database, and disease related genes from the HPO and DrugBank databases.Furthermore, disease association information was mapped to the disease ontologies of the OMIM and Orphanet databases (Fig. 1).

Database Construction

MySQL was employed for data storage, Node.js was utilized to develop the backend service, and the Vue.js framework was adopted to construct the frontend web interface, thereby supporting data browsing, retrieval and visual representation. The specific functional design comprises data storage, data retrieval, and data visualization.

Database Structure Design

TCMCPD was designed with a total of 7 data tables, namely the Patient Table, Medical Record Table, TCM Syndrome Table, TCM Symptom Table, TCM Herb Table, Ingredient Table, and Target Table. The relationships among these entities were established through 5 relationship tables(Fig. 2). Among them, a one-to-many relationship was established between patients and medical records, i.e., each patient may have multiple clinical visits, which generate multiple corresponding medical records. A many-to-many relationship was defined between medical records and TCM syndromes, TCM symptoms, and TCM herbs, which captures the TCM syndromes, manifested TCM symptoms, and prescribed TCM herbs for patients across different clinical visits. Meanwhile, many-to-many relationships were constructed between TCM herbs and ingredients, as well as between ingredients and targets, which reflects that a single TCM herb may contain multiple ingredients, while each ingredient may act on multiple pharmacological targets. Through this relational design, the database can effectively store and manage patients' TCM-related data, ensuring that all types of information can be precisely associated and queried. Meanwhile, the database is equipped with modular data retrieval functions, allowing users to perform exact or fuzzy queries on any field of each data table within the database. The backend uses Node.js to build API interfaces for efficient interaction with the MySQL database, and adopts asynchronous invocation to realize real-time data return.

Direct associations between TCMCPD components

The TCMCPD contains a total of 6 sets of direct relationships, namely Patient-Medical Record, Medical Record-TCM Syndrome, Medical Record-TCM Symptom, Medical Record-TCM Herb, TCM Herb-

Ingredient, and Ingredient-Target (Fig. 2).Among these, based on the structured medical record texts, direct relationships between patients, medical records, TCM syndromes, TCM symptoms, and TCM herbs were constructed in this study. First, the direct relationship between patients and medical records was obtained through the structuring and standardization processing of TCM medical record texts described above. Then, by matching with the information of TCM syndromes, TCM symptoms, and TCM herbs retrieved from the SymMap Database, three sets of direct relationships were obtained, namely Medical Record-TCM Syndrome, Medical Record-TCM Symptom, and Medical Record-TCM Herb. Meanwhile, two sets of direct relationships, namely TCM Herb-Ingredient and Ingredient-Target, were acquired from existing publicly available databases.

Indirect relationships between TCMCPD components

In addition to the 6 sets of direct relationships mentioned above, there are 6 types of indirect relationships, namely Patient-TCM Syndrome, Patient-TCM Symptom, Patient-TCM Herb, TCM Syndrome-TCM Herb, TCM Symptom-TCM Herb, and TCM Herb-Target (Fig. 1).Among these, the Patient-TCM Syndrome, Patient-TCM Symptom, Patient-TCM Herb, TCM Syndrome-TCM Herb, and TCM Symptom-TCM Herb associations were all constructed with the medical record as the intermediate component. These associations are directly derived from clinical practice and have been validated by TCM experts, eliminating the need for filtering with statistical methods, to ensure the integrity of the TCM theoretical system and its clinical utility. Meanwhile, the TCM Herb-Target relationship uses the ingredient as the intermediate component, and was derived through the integration of the three-layer “TCM Herb-Ingredient-Target” network from multiple existing publicly available databases including SymMap, TCMSP, and HIT. The construction of this relationship relies on well-established pharmacological mechanisms of action and has been validated across multiple databases, thus also eliminating the need for screening with statistical methods, to retain the inherent multi-ingredient, multi-target action characteristics of TCM herbs.

Database Physical Design and Security Strategy

In the database design, full consideration was given to data indexing, transaction processing, and regular backup strategies, to ensure the security, stability, and efficient response of data storage. For data indexing, the B-Tree index was adopted[27]. When creating indexes, the principle of “indexing frequently queried columns” was followed, appropriate columns were selected for indexing to avoid redundant indexes[28], and regular maintenance and optimization of indexes were performed; For transaction processing, BEGIN or START TRANSACTION was explicitly executed before the start of a transaction, COMMIT was used to commit transactions, and TRY-CATCH blocks were employed to capture exceptions, ensuring that transactions were properly rolled back. Meanwhile, to avoid data conflicts and concurrency issues caused by concurrent operations, appropriate locking strategies were implemented[29];For the regular backup strategy, a three-level backup system of “full backup + incremental backup + binary log” was adopted[30]: full backup was performed in the early morning every

Sunday, daily incremental backup was executed in the early morning every day, and real-time synchronization of binary logs was implemented. The backup files were compressed and stored in local and off-site cloud storage, with retention periods of 4 weeks, 7 days, and 14 days for full backups, incremental backups, and binary logs respectively, to achieve zero data loss.

Knowledge Graph Construction

The TCMCPD designed was taken as the knowledge source for the knowledge graph, and knowledge extraction was performed directly using the structured data within the database. The core entities, attributes, and inter-entity relationships in the TCM domain were identified and extracted from the existing data, thereby providing a clear semantic foundation for the construction of the knowledge graph. First, in accordance with the database design, core entities were extracted from the 7 data tables; second, the detailed attribute information corresponding to each entity was obtained from each data table; finally, the explicit relationships between entities were directly extracted using the existing foreign keys and association table information in the database.

The Neo4j Desktop software was used for knowledge graph construction, and the ETL TOOL was employed to import the data stored in MySQL. First, the connection was configured by defining the access information for the source database and the target Neo4j instance; second, via configuration files, tables were mapped to graph nodes (e.g., the Patient Table was mapped to Patient nodes), foreign keys and association tables were mapped to graph relationships (e.g., the TCM Herb-Ingredient association table was mapped to TCM Herb-Ingredient relationships); finally, data was imported into Neo4j in batches, generating labels, attributes, and relationships, which formed structured TCM knowledge triples and were stored. Details are shown in Table. 1.

Table 1
Conceptual Hierarchy and Attributes of Each Ontology

Ontology	Instance	Property
Patient	Patient00001 …	Patient ID ，Patient Information
Medical Record	Visit000001 …	Medical Record ID ，Patient ID，Original Case Text，Treatment Principle，Tongue Manifestation，Pulse Manifestation
TCM Syndrome	Lower Origin Deficiency Cold …	Syndrome ID，Syndrome Name，Syndrome Explanation，Syndrome (English)，Syndrome Pinyin
TCM Symptoms	Belching …	Symptom ID，Symptom Name，Symptom Explanation，Symptom Location，Symptom Property
Herb	Aidicha …	Herb ID，Herb Name，Latin Name，Pinyin Name，Herb Name (English) Herb Nature and Flavor，Herb Nature and Flavor(English)，Medicinal Part，Medicinal Part (English)，Category，Category (English)
Ingredient	Anthocyanidin …	Ingredient ID，Ingredient Name，Chemical Formula，Molecular Weight，Oral Bioavailability Score
Target	A2M …	Target ID，Gene Feature，Chromosome，Gene Name，Protein Name

Results

Database Statistics

The seven components in TCMCPD include 5,172 patients, 12,900 Medical records, 233 TCM syndromes, 1,717 TCM symptoms, 499 TCM herbs, 19,595 ingredients, and 5,235 targets, as detailed in Table. 2.The six sets of direct relationships comprise 12,900 Patient–Medical Record relationships, 11,655 Medical Record–TCM Syndrome relationships, 57,764 Medical Record–TCM Symptom relationships, 27,817 Medical Record–TCM Herb relationships, 1,313 TCM Herb–Ingredient relationships, and 5,013 Ingredient–Target relationships.

Table 2
Database Statistics

Component	Source	Quantity
Patient	Extracted from *Essential Medical Case Studies of China's Modern Renowned Traditional Chinese Medicine Practitioners*	5172
Medical Record	Extracted from *Essential Medical Case Studies of China's Modern Renowned Traditional Chinese Medicine Practitioners*	12900
TCM Syndrome	Extracted from *Essential Medical Case Studies of China's Modern Renowned Traditional Chinese Medicine Practitioners*	233
TCM Symptoms	Extracted from the TCMID database and TCMSP database	1717
TCM Herb	Extracted from the Chinese Pharmacopoeia, TCMID database, and TCMSP database	499
Ingredient	Extracted from the TCMID database and TCMSP database	19595
Target	Extracted from the HIT database, HPO database, and DrugBank database	5235

Database Functions

Data Retrieval and Visualization

We provide a user-friendly and intuitive browsing and interactive interface via the web front-end. First, the web homepage presents the overall functions of the system, each data module, and the overall data interrelationships, which incorporates a top navigation bar (with access entries to 7 data modules), a central search box, and a data correlation display panel (Fig. 3a).Herein, users can click the search button on the web homepage, enter query terms on the search page, and execute the search operation. TCMCPD provides a dedicated search box for each functional section, and supports multiple types of search keywords. For instance, when users perform a search for specific TCM clinical case records, TCMCPD supports 6 types of search keywords, including the original text of TCM cases, symptoms, TCM syndromes, TCM herbs, ingredient, targets, and treatment methods, among others, for the convenience of users.

Upon completion of a targeted search operation within TCMCPD, the matching results corresponding to the entered query terms are displayed on a new interface in the form of a summary table (Fig. 3b), where the first column lists the primary key (typically the unique ID) of the corresponding entry in the database for the retrieved section. Users can click the hyperlink embedded in the ID to navigate to the dedicated detailed information interface. In addition, users can navigate and view the full list of items across all 7 constituent functional modules on the browsing page.

On the detailed information page (Fig. 3c), TCMCPD provides descriptive information of the target entry and its intrinsic interrelationships with other functional modules, which are presented through data

visualization and structured tables. The detailed information page consists of three core sections. The first section contains descriptive information, including identifying information such as the unique ID and name, as well as explanatory information such as corresponding definitions and classification categories. The second section is presented in the form of a standardized structured table, which displays the associations between the current functional module and other modules of the database. The third section visualizes the association information described in the second section: the interconnections between each module are represented in the form of nodes and edges, and clicking on a node can display the corresponding descriptive information of the target entry.

Application Case

To demonstrate the practical application utility of the findings from this study, we designed a typical retrieval and analysis workflow: it is assumed that a user intends to retrieve all outpatient consultations involving the prescription of Codonopsis Radix (dangshen) in the outpatient setting, and conduct an in-depth analysis of the key diagnostic information and concomitant medication profiles of thesecases. First, enter “Codonopsis Radix” as the search term in the search box on the homepage and set “Visit” as the search scope, as specifically shown in Fig. 4a clicking the search button retrieves a list of all consultation visits in which Codonopsis Radix(dangshen) was prescribed, including the original full text of the medical case record corresponding to each consultation, as well as the information on the treatment methods, tongue manifestations and pulse manifestations of the corresponding patient in that specific visit, as specifically shown in Fig. 4b.Within the list, the two columns “Visitid” and “Patientid” are configured as clickable hyperlinks to enable further page redirection and information tracing: (1) Clicking on the “Visitid” of a given record will redirect the system to the detailed information page of the corresponding consultation visit. In addition to the existing information including therapeutic methods, tongue manifestations and pulse manifestations, the page also contains the complete list of symptoms and TCM herbs used in this consultation visit, along with their corresponding graphical visualization. Users can further click the link to redirect to the corresponding detailed page, as specifically shown in Fig. 4c.(2) Clicking on the “Patientid” of a given record will redirect the system to the personal information page of the corresponding patient. On this page, users can view the patient’s basic information such as gender and age, as well as a summarized display of all the patient’s historical consultation visits. Users can also click the “VisitID” of each consultation to access the detailed information page of the specific visit, as specifically shown in Fig. 4d.

Discussion

System Function: Multi-dimensional Data Integration, Retrieval and Visualization

From the perspective of the design and implementation of system functions, the core function of TCMCPD is to achieve the systematic integration and correlation analysis of multi-dimensional TCM data. TCMCPD systematically integrates seven categories of entities, namely patients, Medical records,

TCM syndromes, TCM symptoms, TCM Herbs, ingredients, and targets, and establishes a complete data chain of “Patient-Visit-TCM symptom-TCM syndrome-TCM herb-ingredient-target”. spanning from clinical phenotypes to molecular mechanisms. It not only supports single-dimensional information query, but also enables associative retrieval and knowledge navigation. Through the interactive data visualization function of TCMCPD, users can understand the fundamental knowledge of TCM from a macroscopic perspective, and further delve into the microscopic level to explore the mechanism of action of specific chemical constituents of TCM Herbs, realizing a seamless global-to-local knowledge exploration process. At the level of specific functional implementation, based on knowledge graph technology, all entity nodes and their intrinsic interrelationships are visually presented, enabling clinicians to rapidly identify TCM treatment methods and TCM Herbs combinations corresponding to specific TCM symptoms or TCM syndromes. Meanwhile, TCMCPD supports the mining of historical TCM case and the tracking and evaluation of therapeutic outcomes, providing a valuable reference for the standardization and normalization of TCM clinical diagnosis and treatment.

Clinical Application: Improving the Accuracy of TCM Syndrome Differentiation and Treatment and the Specificity of Medication

From the practical application perspective of clinical clinicians, the core value of TCMCPD lies in improving the accuracy of TCM syndrome differentiation and treatment, as well as the targeted specificity of clinical medication regimens. TCM clinical diagnosis and treatment takes TCM syndrome differentiation and treatment as its core principle, which requires clinicians to conduct a comprehensive analysis of the patient's TCM syndromes, TCM symptoms and physical signs, and match the corresponding TCM formulas and TCM Herbs for treatment[31]. However, the traditional TCM clinical practice model is highly dependent on the personal clinical experience of clinicians, with inherent subjectivity and uncertainty[32].TCMCPD systematically integrates structured TCM case data and TCM Herbs information, providing clinicians with objective reference during the clinical decision-making process.In practical clinical application scenarios, after clinicians complete TCM syndrome differentiation, they can query the high-frequency TCM Herbs for this TCM syndrome or corresponding symptoms through the TCMCPD, and evaluate the potential therapeutic efficacy of these components in specific diseases in combination with modern pharmacological research data. This diagnosis and treatment model helps optimize the selection of TCM Herbs, frees clinicians from the limitations of their own clinical experience, and thereby improves the scientificity and repeatability of clinical therapeutic effects. Meanwhile, TCMCPD supports rapid retrieval of similar historical TCM cases, enabling clinicians to refer to previous successful treatment cases and optimize the treatment regimen for the current patient.

Scientific Research Support: Uncovering the Mechanism of Action and Mining the Compatibility Rules of TCM Herbs

In the field of TCM Herbs research, TCMCPD provides a high-quality data source for uncovering the mechanism of action of TCM Herbs, mining their compatibility rules, and supporting the modern

research and development(R&D) of TCM Herbs.The core challenge faced by traditional TCM Herbs research lies in how to effectively bridge the holistic view of TCM with the molecular mechanisms of modern pharmacology[33].TCMCPD establishes an association path of “TCM Herbs-Ingredients-Targets”, enabling researchers to track the high-frequency TCM Herbs and their main ingredients corresponding to a specific symptom or TCM syndrome, and then analyze the biological targets of these ingredients to uncover the mechanism of action of TCM Herbs at the molecular level. In terms of compatibility rule mining, TCMCPD supports the systematic analysis of the relationships between different TCM Herbs prescriptions, TCM symptoms and TCM syndromes in historical TCM cases, and can identify the compatibility rules and combination patterns of TCM Herbs. In terms of supporting TCM Herbs R&D, TCMCPD can provide data support for the clinical screening and optimization of new drugs. By analyzing high-frequency TCM Herbs and their active ingredients corresponding to specific disease syndromes, researchers can identify candidate drugs with potential therapeutic value and shorten the R&D cycle of new TCM Herbs. In addition, the TCMCPD can also provide a scientific reference for the precise application of TCM Herbs and the formulation of individualized treatment regimens.

AI Integration: Empowering Intelligent Question Answering for TCM Knowledge and Decision Support for TCM Practice

From the perspective of AI application, TCMCPD, as a high-quality structured knowledge base, provides a solid data foundation for the application of AI technology in the field of TCM, and supports the training and optimization of AI models. It not only enables the models to automatically identify TCM syndromes, extract key symptoms and deduce the compatibility rules of TCM Herbs, but also helps the models better understand the logic of TCM syndrome differentiation and treatment, as well as the active ingredients and mechanism of action of modern medicines, thus promoting the innovative development of the integrated TCM and Western medicine diagnosis and treatment model. In terms of the integration of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) technology, since LLMs are mainly trained on general historical data, they carry an inherent risk of “hallucinations”[34] in vertical fields such as TCM. The structured nature of knowledge graphs enables the introduction of external information sources via RAG technology, which effectively prevents the generation of erroneous information[35]. Therefore, integrating the knowledge graph within TCMCPD with LLMs enables the construction of an intelligent TCM question-and-answer (Q&A) system[36, 37]. This system can convert questions raised by users into structured query requests to ensure the accuracy of information[38], and can further provide clinical decision support and individualized diagnosis and treatment recommendations, delivering scientifically rigorous and highly instructive treatment regimens for clinicians and patients.

Education Upgrading: TCM Case Inheritance Education Platform and Integrated TCM-Western Medicine Thinking Cultivation

For a long time, traditional TCM education has followed the oral instruction and personal impartation model of “master-apprentice mentoring”, which is plagued by inherent problems such as low efficiency of knowledge inheritance, scattered learning resources, and difficulties in accessing clinical experience[39, 40]. TCMCPD systematically integrates multi-dimensional data including patients, TCM cases, TCM syndromes, TCM symptoms, TCM Herbs, ingredients and targets, thus providing a standardized and structured teaching resource platform for TCM education, and helping to promote the transformation of the TCM education model from experience-based inheritance to data-driven modernization. In terms of teaching resource development, the historical TCM cases in TCMCPD can serve as a typical case library for students' learning. Through the visual interface of the knowledge graph, students can learn the diagnosis and treatment experience of renowned veteran TCM physicians from a macro perspective, then gradually delve into the details of specific TCM syndrome differentiation and prescription compatibility, thus realizing systematic learning. Meanwhile, the “TCM Herbs-ingredients-targets” association path in the database enables students to understand the molecular mechanisms of modern pharmacology of prescriptions while learning traditional TCM theories, which promotes the cultivation of integrated TCM and Western medicine thinking.

Conclusions

In this study, we constructed a TCM case database covering the complete relational chain of “Patient-Visit-TCM symptom-TCM syndrome-TCM herb-ingredient-target”.. It not only integrates the basic information and diagnosis and treatment records of TCM cases, but also incorporates the chemical ingredients and target data of TCM Herbs from existing TCM databases. For clinicians, it provides structured TCM case data and references for high-frequency TCM Herbs application, so as to improve the accuracy of TCM syndrome differentiation and treatment, enhance the specificity of medication, and reduce the uncertainty of empirical medication. For researchers, it supports the revelation of the mechanism of action of TCM Herbs, the mining of compatibility rules, and R&D of new TCM Herbs. For AI applications, it serves as a high-quality structured knowledge base to support the construction of intelligent TCM Q&A and clinical decision support systems. For TCM education, it provides a professional teaching resource platform to promote the innovative development of TCM education and the cultivation of integrated TCM and Western medicine thinking. In summary, TCMCPD has realized multi-dimensional value application covering clinical practice, scientific research, intelligent application, and medical education.

TCMCPD also has certain limitations. The database is mainly constructed based on *Essence of Medical Cases of Modern Famous TCM Physicians in China*, with a single data source. Ancient TCM cases, electronic medical records (EMRs) from modern TCM hospitals, and other types of TCM case data have not been effectively integrated. Thus, its representativeness and coverage fail to fully reflect the diversity and regional differences of TCM cases. In the future, we will further expand data sources, include TCM cases from more regions and academic schools of TCM, improve the database management system, expand front-end web page functions, and implement dynamic update and real-time maintenance of the database content.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Institute of Management, Beijing University of Chinese Medicine, Beijing 102488, China.

[2] Department of Information Management, Peking University, Beijing 100871

Funding

Not applicable.

Author Contribution

GFY designed the study. CJB participated in data collection and database construction. MY provided technical support and performed data analysis. ZBT and WRJ provided critical academic advice for the study. CJB, MY and CG drafted the manuscript. All authors have read and approved the final version of the manuscript.

Acknowledgement

This work was supported by the following projects: 2025 Beijing Municipal Social Science Fund Planning Project: Research on the Allocation and Governance Mechanism of Information Resources in Online Healthcare Communities from the Multi-stakeholder Perspective (Grant No. 25BJ03238);Chaoyang District Digital Healthcare Proof of Concept (PoC) Program Project: “TCM-featured Community Full-cycle Intelligent Agent for Health Maintenance, Elderly Care, Health Education and Medical Services” (Project No. 2025SLQY007);Research and Practice on the Construction of the General Education Course “Introduction to Artificial Intelligence” for TCM Universities and Colleges, Open Bidding and Chief-led Teaching Reform Project of Beijing University of Chinese Medicine (Grant No. JX-JBGS-2501005).

Data Availability

Main data and related functions are publicly available through a web interface at http://182.92.96.53:8081/. To obtain the complete data, please contact the corresponding author.

References

1. Zhou X, Peng Y, Liu B. Text mining for traditional Chinese medical knowledge discovery: a survey. J Biomed Inf. 2010;43(4):650–60. https://doi.org/10.1016/j.jbi.2010.01.002.
2. Zhou E, Shen Q, Hou Y. Integrating artificial intelligence into the modernization of traditional Chinese medicine industry: a review. Front Pharmacol. 2024;15:1181183. https://doi.org/10.3389/fphar.2024.1181183.
3. Zhai Y, Liu L, Zhang F, et al. Network pharmacology: a crucial approach in traditional Chinese medicine research. Chin Med. 2025;20(1):8. https://doi.org/10.1186/s13020-024-01056-z.
4. Chen H, Tang R, Hong M, et al. Applications of artificial intelligence in the research of molecular mechanisms of traditional Chinese medicine formulas. Chin J Nat Med. 2025;23(11):1329–41. https://doi.org/10.1016/S1875-5364(25)60903-4.
5. Liu B, Zhou X, Wang Y, Hu J, He L, Zhang R, Chen S, Guo Y. Data processing and analysis in real-world traditional Chinese medicine clinical data: challenges and approaches. Stat Med. 2012;31(7):653–60. https://doi.org/10.1002/sim.4417.
6. Yao L, Jin Z, Mao C, Zhang Y, Luo Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. J Am Med Inf Assoc. 2019;26(12):1632–6. https://doi.org/10.1093/jamia/ocz164.
7. You M, Chen C, Li GZ, et al. ISMAC: An Intelligent System for Customized Clinical Case Management and Analysis. ScientificWorldJournal. 2015;2015:473168. https://doi.org/10.1155/2015/473168.
8. TCM Case Think Tank Platform. https://www.jinjuya.com/.Accessed 11 April 2026.
9. Ancient and Modern TCM Case Cloud Platform. https://www.yiankb.com/.Accessed 11 April 2026.
10. Iwakabe S, Gazzola N. From single-case studies to practice-based knowledge: aggregating and synthesizing case studies. Psychother Res. 2009;19(4–5):601–11. https://doi.org/10.1080/10503300802688494.
11. Zhang L, Jin Y, Li J, et al. Epidemiological research on rare diseases using large-scale online search queries and reported case data. Orphanet J Rare Dis. 2023;18(1):236. https://doi.org/10.1186/s13023-023-02839-7.
12. Deng SZ, Liu XY, Su JJ, et al. Epidemiological and cluster characteristics of dengue fever in Yunnan Province, Southwestern China, 2013–2023. BMC Infect Dis. 2025;25(1):95. https://doi.org/10.1186/S12879-024-10403-2.
13. Zack T, Dhaliwal G, Geha R, Margaretten M, Murray S, Hong JC. A Clinical Reasoning-Encoded Case Library Developed through Natural Language Processing. J Gen Intern Med. 2023;38(1):5–11.

https://doi.org/10.1007/S11606-022-07758-0.

14. Zhai Y, Liu L, Zhang F, et al. Network pharmacology: a crucial approach in traditional Chinese medicine research. Chin Med. 2025;20:8. https://doi.org/10.1186/s13020-024-01056-z.
15. Zhang RZ, Yu SJ, Bai H, Ning K. TCM-Mesh: The database and analytical system for network pharmacology analysis for TCM preparations. Sci Rep. 2017;7(1):2821. https://doi.org/10.1038/s41598-017-03039-7.
16. Ru J, Li P, Wang J, et al. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. J Cheminform. 2014;6:13. https://doi.org/10.1186/1758-2946-6-13.
17. Ye H, Ye L, Kang H, et al. HIT: linking herbal active ingredients to targets. Nucleic Acids Res. 2011;39(Database issue). https://doi.org/10.1093/nar/gkq1165. D1055-9.
18. Chen CY. TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. PLoS ONE. 2011;6(1):e15939. https://doi.org/10.1371/journal.pone.0015939.
19. Peri S, Navarro JD, Amanchy R, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 2003;13(10):2363–71. https://doi.org/10.1101/GR.1680803.
20. Zhang R, Zhu X, Bai H, Ning K. Network Pharmacology Databases for Traditional Chinese Medicine: Review and Assessment. Front Pharmacol. 2019;10:123. https://doi.org/10.3389/fphar.2019.00123.
21. Chen J, Guo F, Zhao B, et al. Research on multi-dimensional named entity recognition in traditional Chinese medicine cases. Chin Arch Tradit Chin Med. 2026;44(2):17–23. https://doi.org/10.13193/j.issn.1673-7717.2026.02.004. (in Chinese).
22. Lample G, Ballesteros M, Subramanian S et al. Neural architectures for named entity recognition. Proc Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol. 2016:260–70. https://doi.org/10.48550/arXiv.1603.01360
23. Strubell E, Verdie V, Gibson E et al. Fast and accurate entity recognition with iterated dilated convolutions. Proc Conf Empir Methods Nat Lang Process. 2017:2670–80. https://doi.org/10.48550/arXiv.1702.02098
24. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008. https://doi.org/10.48550/arXiv.1706.03762.
25. Dai Z, Wang X, Ni P et al. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. Proc 12th Int Congr Image Signal Process BioMed Eng Inform (CISP-BMEI). 2019:1–5. https://doi.org/10.1109/CISP-BMEI48845.2019.8965823
26. Wu Y, Zhang F, Yang K, et al. SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. Nucleic Acids Res. 2019;47(D1):D1110–7. https://doi.org/10.1093/nar/gky1021.
27. Yu T, Zou Z, Sun W, Yan Y. Refactoring index tuning process with benefit estimation. Proc VLDB Endow. 2024;17(7):1528–41. https://doi.org/10.14778/3654621.3654622.

28. Kossmann J, Halfpap S, Jankrift M, Schlosser R. Magic mirror in my hand, which is the best in the land? An experimental evaluation of index selection algorithms. Proc VLDB Endow. 2020;13(11):2382–95. https://doi.org/10.14778/3407790.3407832.
29. Elhoseny M et al. An efficient deadlock handling model based on neutrosophic logic: case study on real time healthcare database systems. IEEE Access. 2022;10:98328-98342.10.https://doi.org/1109/ACCESS.2022.3192414.
30. Taft R, Sharif I, Matei A et al. CockroachDB: The resilient geo-distributed SQL database. Proc ACM SIGMOD Int Conf Manag Data. 2020:1493–509. https://doi.org/10.1145/3318464.3386134
31. Gong L, Jiang J, Li X, Guo Y. A syndrome differentiation model of TCM based on multi-label deep forest using biomedical text mining. Front Genet. 2023;14:1272016. https://doi.org/10.3389/fgene.2023.1272016.
32. Yan JY, Zhang PW, Sheng WG, et al. Evolutionary fuzzy learning for Chinese medicine liver syndrome differentiation. Syst Sci Control Eng. 2025;13(1):2507261. https://doi.org/10.1080/21642583.2025.2507261.
33. Ren Z, Ren Y, Li Z, Xu H. TCMM: A unified database for traditional Chinese medicine modernization and therapeutic innovations. Comput Struct Biotechnol J. 2024;23:1619–30. https://doi.org/10.1016/j.csbj.2024.04.016.
34. Wang B, Lu Y, Wang Z, Ge P, Wang G, Yao K, Peng S, Zhu Y. Sci Data. 2025;13(1):79. https://doi.org/10.1038/s41597-025-06387-6. TCMEval-PA: a question-answering benchmark dataset for the prescription audit of Traditional Chinese Medicine.
35. Evangelista E, Ruba F, Bukhari S, Nazir A, Sharma R. GraphRAG-enabled local large language model for gestational diabetes mellitus: development of a proof-of-concept. JMIR Diabetes. 2026;11:e76454. https://doi.org/10.2196/76454.
36. Liu C, Chang Y, Li J, et al. Improving TCM question answering through tree-organized self-reflective retrieval with LLMs. Front Med. 2026;13:1752778. https://doi.org/10.48550/arXiv.2502.09156.
37. Chen J, Chen J, Luo G, et al. Construction and evaluation of the knowledge graph and large model question-answering system for Jin San Zhen therapy: a tool study for primary care and general practice. Front Med. 2026;13:1755583. https://doi.org/10.3389/fmed.2026.1755583.
38. Liu F, Liu M, Li M, et al. Automatic knowledge extraction from Chinese electronic medical records and rheumatoid arthritis knowledge graph construction. Quant Imaging Med Surg. 2023;13(6):3873–90. https://doi.org/10.21037/qims-22-1158.
39. Xiang S, Lin H, Cai F, Jiang Z, et al. Integrating knowledge graphs with ancient Chinese medicine classics: challenges and future prospects of multi-agent system convergence. Chin Med. 2025;20(1):168. https://doi.org/10.1186/s13020-025-01226-7.
40. Yang F, Yang T, Li Y, Wang D. The evolution of traditional Chinese medicine education policies in China: themes, characteristics, and experiences—a historical institutionalism analysis based on 197 policy texts. Front Public Health. 2026;13:1692198. https://doi.org/10.3389/fpubh.2025.1692198.

Figures

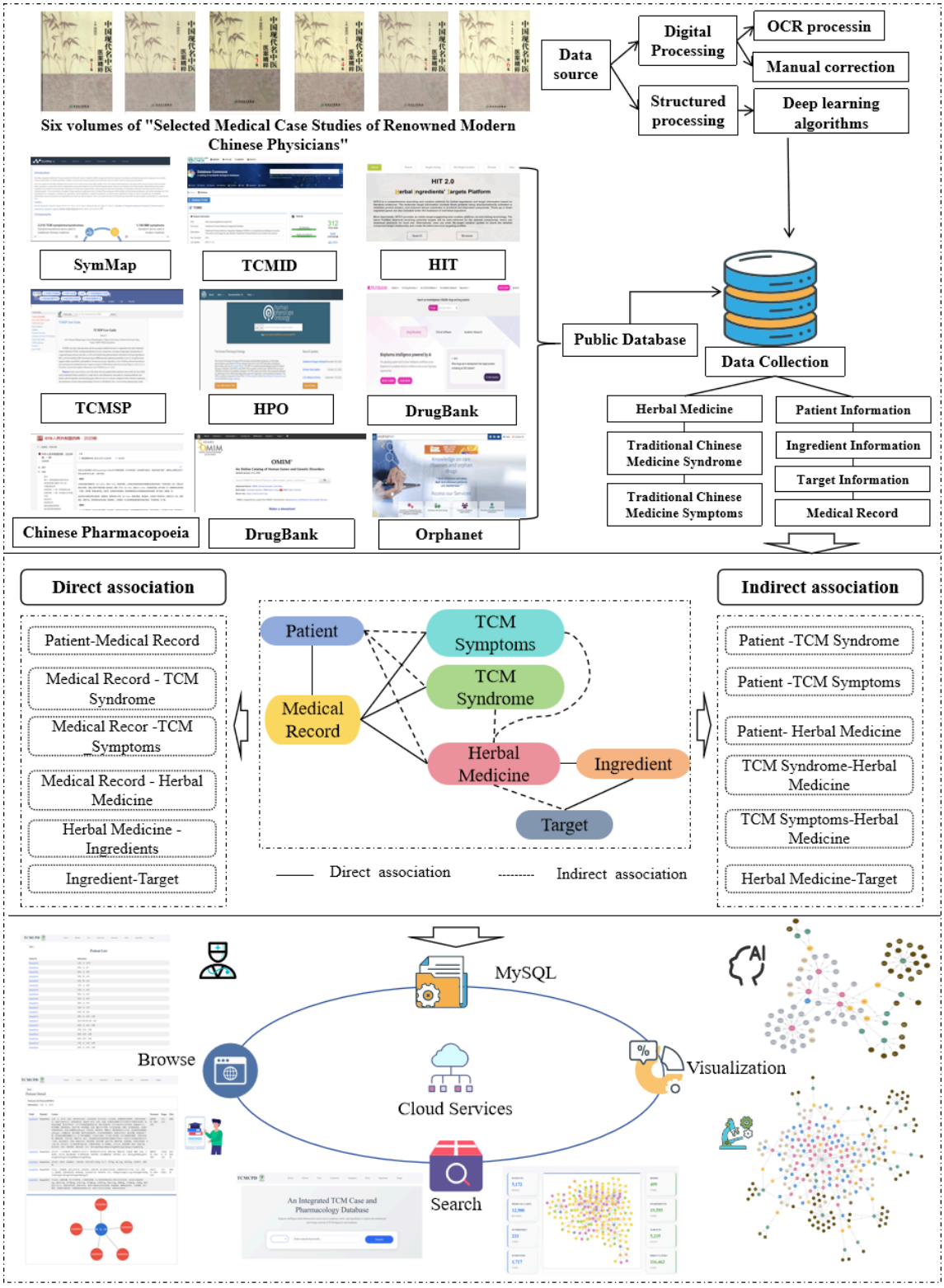


Figure 1

Database construction architecture diagram

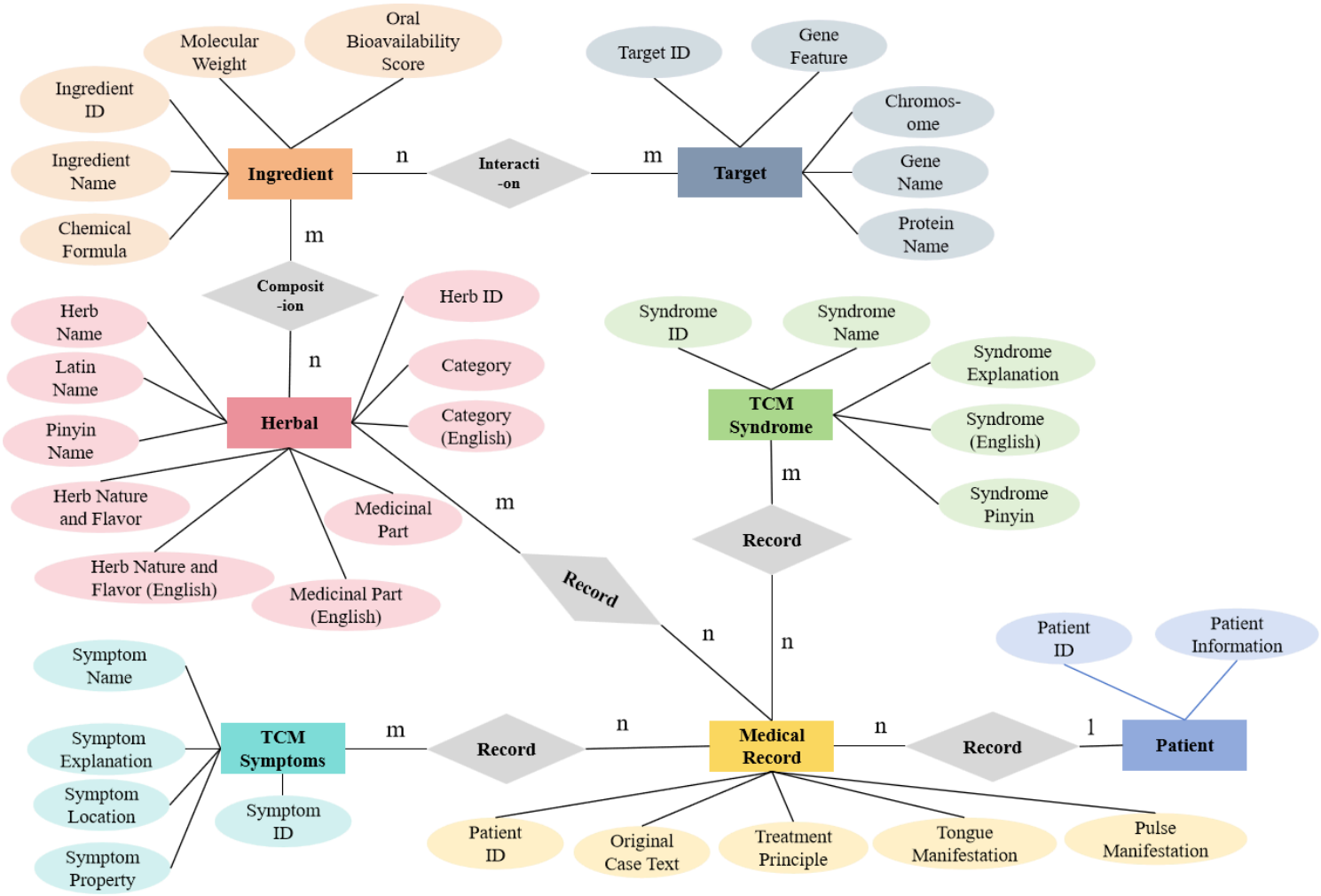


Figure 2

Entity-Relationship Diagram (E-R Diagram)

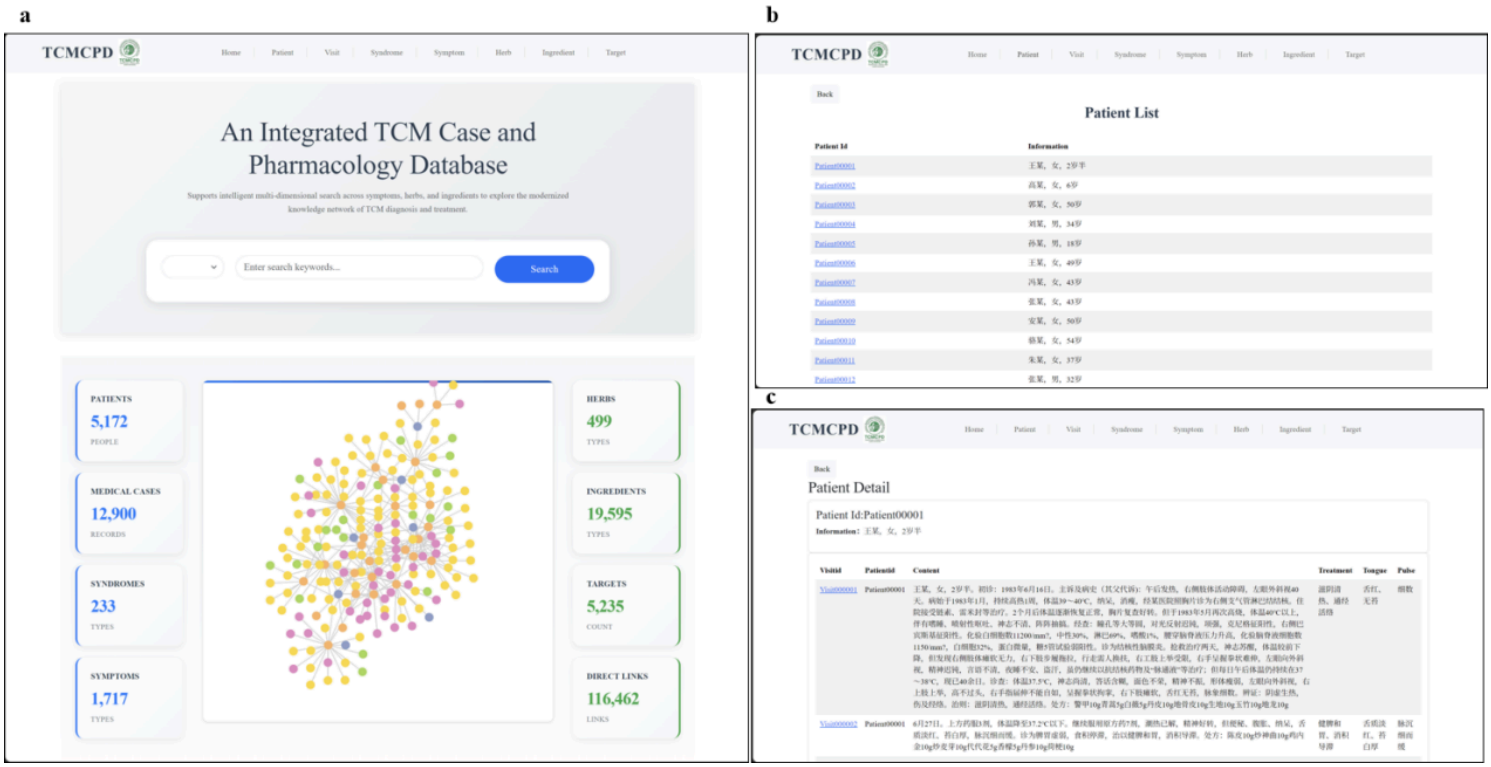


Figure 3

Schematic Diagram of the TCMCPD Search Function. **a** TCMCPD Primary-Level Page: Homepage. **b** TCMCPD Secondary-Level Page: Search and List Display Page. **c** TCMCPD Tertiary-Level Page: Detailed Information Display Page.

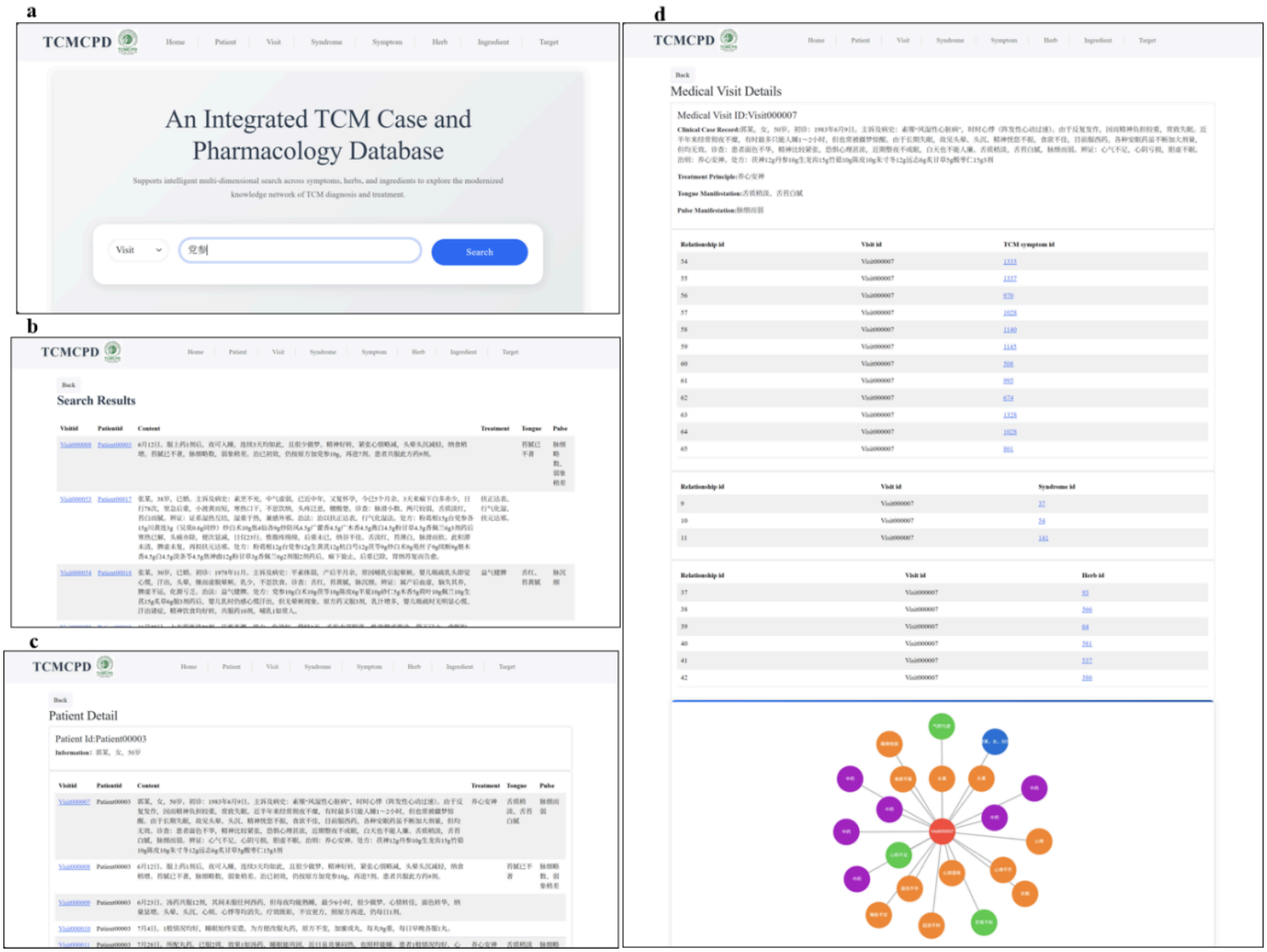


Figure 4

Application Cases. **a** TCMCPD “Codonopsis Radix” Search Page. **b** TCMCPD Search Results Page. **c** TCMCPD Patient Detailed Information Page. **d** TCMCPD Medical Case Record Detailed Information Page