Methodology and Empirical Study for Constructing Traditional Chinese Medicine Diagnosis and Treatment Datasets Based on Medical Case Metadata (Postprint)

Authors: Ma Jie, Wang Jue, Sun Hengyu, Li Lu

Date: 2023-04-01T16:02:44+00:00

Abstract

[Objective/Significance] TCM case records are an important carrier for documenting clinical data. Collecting TCM clinical data based on TCM case record metadata to form datasets is of great significance for the sharing, mining, and inheritance of TCM clinical knowledge. [Methods/Process] This study selects and determines TCM case record metadata, conducts conceptual and logical design of the dataset with reference to relational databases, proposes a method for collecting, organizing, and storing TCM case record data to form datasets, and performs empirical research. [Results/Conclusion] Based on relatively complete TCM case record metadata for online and on-site clinical data collection, the resulting TCM clinical dataset can not only support multi-dimensional storage, sharing, and retrieval of clinical information, but also provide a data source for more in-depth clinical data mining.

Full Text

Construction Methods and Empirical Research of Traditional Chinese Medicine Diagnosis and Treatment Dataset Based on Medical Case Metadata

Ma Jie[1],[2], **Wang Jue**[1], **Sun Hengyu**[3], **Li Lu**[1] [1]School of Management, Jilin University, Changchun 130022 [2]Information Resources Research Center of Jilin University, Changchun 130022 [3]The First Clinical Hospital of Jilin Academy of Traditional Chinese Medicine, Changchun 130022

Abstract: [Purpose/Significance] Traditional Chinese Medicine (TCM) medical cases are important carriers for recording diagnosis and treatment data. Collecting TCM diagnosis and treatment data based on TCM medical case

metadata to form a dataset is of great significance for the sharing, mining, and inheritance of TCM diagnosis and treatment knowledge. [Method/Process] We selected and determined TCM medical case metadata, referenced relational databases for conceptual and logical design of the dataset, proposed methods for collecting, organizing, and storing TCM medical case data to form a dataset, and conducted empirical research. [Result/Conclusion] Based on relatively complete TCM medical case metadata, we collected data from both online and field diagnosis and treatment. The resulting TCM diagnosis and treatment dataset can not only support multi-dimensional storage, sharing, and retrieval of diagnosis and treatment information, but also provide data sources for deeper diagnosis and treatment data mining.

Keywords: medical case; TCM medical case metadata; TCM diagnosis and treatment dataset

---

After millennia of development and accumulation, Traditional Chinese Medicine has formed its own unique mode of thinking. However, its dissemination and development still rely primarily on "oral teaching and personal example." Due to TCM's inherent characteristics of being "patient-centered" and prescribing "different formulas for different individuals," a widely adopted diagnostic and treatment model has not been established. Consequently, a large amount of high-value data from TCM diagnosis and treatment cannot be effectively preserved and analyzed, limiting the development of TCM. The root cause lies mainly in poor information dissemination and communication barriers. The establishment of domain-specific datasets and the publication of high-quality scientific data are important means to achieve data sharing and break information silos. Therefore, proposing a set of scientifically rigorous metadata for recording TCM diagnosis and treatment details, and establishing a TCM diagnosis and treatment dataset based on this metadata to document and preserve the vast amount of data generated during TCM diagnosis and treatment processes for data analysis and tacit knowledge mining, can greatly promote the development and utilization of TCM diagnosis and treatment information.

Currently, most TCM medical cases are recorded in unstructured narrative text, with coarse field granularity that fails to reveal the details of TCM cases, making knowledge mining difficult. Therefore, structuring unstructured text is essential. First, granularity should be refined by developing appropriate metadata to structure important information from case texts. Second, terminology standardization should be performed by mapping medical terms (such as drug names and disease names) in the text to authoritative or standardized forms. This not only reduces defects and errors in treatment records and improves data quality, but also facilitates the sharing and exchange of professional knowledge. In summary, this paper proposes constructing a high-quality TCM diagnosis and treatment dataset based on TCM medical case metadata to achieve structured storage of TCM diagnosis and treatment data and support deep retrieval and data mining.

1 Literature Review

TCM diagnosis and treatment involves complex and diverse knowledge. To ensure dataset quality, a reasonable collection domain should be selected. As Zhang Taiyan stated, "The most outstanding achievements of TCM are found in medical cases." TCM medical cases represent the comprehensive application of TCM principles, methods, formulas, and medicines. They not only faithfully document medical activities but also reflect physicians' clinical experience and diagnostic thinking. Most importantly, they record detailed patient information rather than focusing solely on the disease itself. Therefore, historical medical cases, online TCM cases, and clinically recorded cases all serve as data sources for dataset collection. However, regardless of the type, these records are generally unstructured narrative texts with coarse field granularity that cannot reveal the details of TCM cases, making knowledge mining difficult.

Since the 1990s, different metadata standards have emerged in various application domains. To standardize information storage and facilitate resource sharing and utilization, scientific data organization in relevant fields both domestically and internationally has mostly adopted metadata approaches. However, there are numerous metadata models, and even within the same field, no universally recognized metadata model can satisfy all applications. According to Jia Lirong et al.'s research on ISO metadata standards, there is currently no complete and widely used metadata system for the TCM field, requiring extension of existing metadata based on specific use cases. Zhao Yang et al., through analysis of TCM literature metadata objects and necessity, proposed a framework structure and extension principles for TCM literature metadata, providing ideas for metadata development in the TCM field. Liu Baojie et al. designed a metadata scheme for TCM medical record databases based on extensive review of cases, dividing it into nine components including basic case information, disease information, diagnosis, treatment, prescriptions, errors, contraindications, Western medicine information, and medical orders. While this achieved preliminary organization and description of TCM case resources, it suffered from being "comprehensive but not refined," with overly general field definitions and lack of relevant standards, causing misalignment and confusion in data collection. Sun Jing et al., based on ontology theory, explored information acquisition and management methods for TCM medical cases. Using ontology to organize data can clearly reveal intrinsic relationships between data, largely compensating for metadata's shortcomings in knowledge organization and facilitating the construction of knowledge sharing systems. However, this research's mining of TCM cases was relatively superficial, with coarse field granularity that limited clinical application.

A dataset is a collection of data with a specific theme that can be identified and processed by computers. It represents a new data organization method and "encapsulation" unit for collecting and organizing data resources based on themes. Constructing domain-specific datasets is currently a widely adopted method for collecting, integrating, and retrieving data resources. The main approaches in-

clude selecting data from domain literature or conducting structured collection of relevant thematic data. For example, Shan Lianhui et al. collected and integrated relevant data from PubMed and SCIE databases to construct a medical field literature evaluation dataset based on DOI, though this method is affected by database and search term selection, with its accuracy and comprehensiveness only preliminarily verified, and its application to other medical domains requires further research. Qiu Ruijin et al. formed a core dataset for clinical safety evaluation through literature research and semi-structured interviews, using Delphi surveys to revise the data, ultimately forming a core dataset through consensus meetings. While this method considered as many relevant factors as possible, complete consensus on core data selection may be unattainable. For medical data, the scope of related data is extensive, involving disease treatment, public health, demographics, and other aspects. Therefore, constructing medical datasets requires reasonable limitation of application scenarios and thematic scope to avoid data explosion from excessive comprehensiveness. Additionally, subjective limitations of constructors and domain experts must be considered, with interviews and Delphi surveys commonly used for data selection and revision.

Currently, established medical datasets both domestically and internationally play important supporting and auxiliary decision-making roles in health management, medical imaging management, and disease prediction. Globally, numerous medical case analysis datasets have been established, such as OASIS and NSCLC, as well as medical denoising and segmentation datasets, most of which are image datasets like brain structure datasets and breast datasets. For recording clinical information, the MIMIC dataset is notable, documenting patient clinical information including demographics, bedside vital signs measurements, nursing notes, discharge information, and related imaging reports. MIMIC supports various analytical studies covering epidemiology, clinical decision rule improvement, and electronic tool development. This dataset is massive, containing almost all static data a patient might involve. Although its recording method for diagnosis and treatment is more suitable for Western medicine, its overall structure provides valuable reference for this TCM dataset construction.

To promote resource sharing in the TCM field, China has established multiple knowledge bases for TCM data management and research, such as the CNKI TCM Knowledge Resource Database, Disease Diagnosis and Treatment Knowledge Base, and Wanfang Medical Network Clinical Diagnosis and Treatment Knowledge Base, which contain rich resources on diseases, examinations, drugs, evidence-based medicine, cases, and literature to provide reliable support for clinical decision-making. A standard knowledge base should have inference mechanisms enabling knowledge reasoning and mining. However, some knowledge bases suffer from chaotic storage in the factual data layer, lack of reasonable table structures, and unscientific metadata field division that fails to reflect TCM characteristics, resulting in data redundancy and conflicts. Their functions remain at the database level, essentially being collections of data from different domains. Whether constructing TCM case datasets or knowledge bases,

the applied metadata must fully structure valuable diagnostic and treatment information while performing terminology standardization and normalization.

2 TCM Medical Case Metadata Selection and Terminology Standardization

2.1 Metadata Selection

Metadata is structured data that describes data attributes, primarily functioning in resource organization, mining, interoperability, digital identification, and preservation. If most clinical data remain scattered and disordered, with non-uniform recording formats and no standardized archiving methods, resources cannot be effectively shared and retrieved, causing massive loss of clinical data and forming information silos. Currently, there is no internationally recognized metadata standard for TCM medical records. Due to the linguistic characteristics of TCM, although internationally 通用 metadata standards have good interoperability, they cannot accurately describe TCM resources. The free narrative style of TCM case records also limits metadata standard selection.

Our research team member Li Lu et al. compiled and designed a set of TCM medical case metadata, which was refined through comprehensive analysis of numerous standards and literature using content analysis, expert consultation, and field research methods, resulting in a relatively reasonable, universally applicable, and highly precise TCM medical case metadata scheme. This dataset construction uses this metadata scheme as a blueprint, with minor adjustments to the original metadata structure to better adapt to data collection and preservation.

2.2 Metadata Structure Revision

To enable medical workers to input data conveniently and efficiently, allow dataset managers to screen and organize data within the dataset, and ensure exported data quality meets researchers' analytical requirements, the overall metadata is divided into two major structures based on function and characteristics: medical case descriptive metadata and medical case management metadata. Medical case descriptive metadata details the diagnosis and treatment process, while medical case management metadata objectively records the preservation and circulation of the cases themselves.

Analysis of existing online medical cases reveals several situations: (1) Basic patient information is often omitted for privacy protection; (2) Patient chief complaints are often described together with the four diagnostic methods, where patients state their main conditions followed by physicians' comprehensive analysis through inspection, auscultation & olfaction, inquiry, and palpation; (3) Physician diagnosis and prescription medical orders are closely connected. Therefore, balancing information volume across sections and considering the overall context and internal logic of medical cases, the medical case descriptive metadata

portion is divided into five subsets: case identification information, patient basic information, diagnosis and treatment process information, diagnosis and treatment method information, and imaging examination information. From the perspective of case preservation management, referencing current hospital outpatient medical record storage rules, a case management and circulation information subset is established to record specific information about case preservation and transfer. The specific structure is shown in [Figure 1: see original paper].

After clarifying the metadata structure, simple field division is performed for each information table to prepare for subsequent design, as summarized in .

2.3 Terminology Standardization

In the various information subsets above, due to differences in data content, sources, and purposes, different standards are adopted within each subset. Standardized terminology is the foundation of medical informatics development, providing a basic framework for recording information, ensuring consistency of original data, and serving as an effective means to solve semantic interoperability, knowledge expression consistency, and medical resource sharing. In the TCM field, the "Semantic Network Framework for Traditional Chinese Medicine Language System" specifies semantic types, concepts, and relationships for TCM language systems with detailed definitions. This standard not only regulates and supports the construction of TCM language systems but also provides semantic standards for TCM terminology systems and ontology creation, supporting mapping for TCM language systems and facilitating exchange of TCM terminology information.

According to medical field terminology standardization requirements, medical terms should use those published by the National Committee for Terms in Sciences and Technologies. For terms without established standards, refer to :

** Standards for Uncertain Medical Terms** | Subject Thesaurus and Terminology System | Application Scope | Description | |————|————|————-| | Medical Subject Headings (MeSH) | Unapproved terms | Standardized thesaurus for indexing, cataloging, and online retrieval | | Chinese Traditional Medicine Subject Headings | Unapproved terms | Establishes foundation for TCM scientific information retrieval system | | TCM Clinical Diagnosis and Treatment Terminology | TCM terms | Standardizes wording for TCM medical records, including disease, syndrome, and treatment method sections | | Pharmacopoeia of the People's Republic of China | Terms without 通用 translation | Legal drug names in China, published by National Pharmacopoeia Commission | | China Approved Drug Names | Terms without 通用 translation | Official drug names formulated by National Pharmacopoeia Commission | | Acupoint Locations, Auricular Point Names | Meridian and acupuncture terms | Recording method references WHO International Standard Acupuncture Point Names |

3 TCM Diagnosis and Treatment Dataset Construction Methods

3.1 Dataset Construction Process

Considering comprehensive dataset applications, several issues should be noted during construction: (1) The dataset's primary function is to record detailed diagnosis and treatment processes, requiring numerous metadata with reasonable field division, concise and clear content, and clear logical hierarchy; (2) To better form a data sharing ecosystem, the dataset should be structurally marked to achieve data co-construction, sharing, and knowledge exchange; (3) Since one application scenario is field data collection, dataset collection interface design and subsequent data management system development are also important considerations for dataset improvement and maintenance. Combining preliminary preparation work, the overall dataset construction process is shown in [Figure 2: see original paper].

3.2 Dataset Structure Design

Dataset structure is similar to relational database structure, comprising a hierarchical structure of tables, rows, columns, and other objects, while also including constraints and relationships defined by the dataset. Therefore, this study references relational database construction methods and steps for dataset structure design.

3.2.1 Dataset Conceptual Structure Design Database construction follows three normal forms that provide basic logical requirements for database design, with appropriate modifications made based on specific circumstances. The first normal form states that "each attribute in an entity tuple is indivisible." Most columns in the above design satisfy this indivisible attribute requirement. However, the "auxiliary examination" field in the patient basic information subset is extremely complex and voluminous, as patients undergo numerous examination methods, each with specialized terminology. For simplicity in data entry and convenience in data retrieval, exhaustive enumeration is not performed here, and all auxiliary examination information is merged into a single column. Similarly, in the "treatment method" field of the diagnosis and treatment method information subset, each drug's name, dosage, usage method, and precautions have one-to-one correspondences and should not be extracted separately, as this would diminish data value. Therefore, metadata such as "prescription drug names" are merged here.

The second normal form states that "no attribute unrelated to the primary key should exist in an entity tuple." Considering privacy protection, fields like "name" involving personal privacy are often omitted from medical cases and thus are unsuitable as primary keys. Therefore, a "case number" field is added to each data table as the primary key, placing each table in proper form.

The third normal form states that "non-key attributes in entity tuples should have no dependency relationships." Each column in every table is independent, with division conforming to basic medical case logic, avoiding data redundancy and information confusion.

Based on these rules, various data tables are created, and data constraints are applied according to actual conditions and requirements:

(1) Case Identification Information Table (): Records the case number, title, source, creator/physician name, patient visit date, case collection date, TCM category, and disease-syndrome classification.

(2) Patient Basic Information Table (): Detailed patient personal information including name, gender, and other basic information, as well as past medical history and personal lifestyle history that may affect conditions. Data standards reference WS445.11-2014 (Part 11 of the People's Republic of China Health Industry Standard - TCM Inpatient Medical Record Front Page), which lists GB/T2261-2003 (Personal Basic Information Classification and Codes) and other standards.

(3) Diagnosis and Treatment Process Information Table (): Records the clinical disease diagnosis process. Traditional TCM understands conditions through the four diagnostic methods (inspection, auscultation & olfaction, inquiry, and palpation), supplemented by patient chief complaints. Physicians comprehensively evaluate current symptoms and disease progression to determine conditions. These subjective descriptions of conditions are highly flexible, requiring physicians to transform patient complaints into medical information through clinical experience. Recording primarily references GB/T15657-1995 (TCM Disease-Syndrome Classification and Codes) and GB/T16751.2-1997 (TCM Clinical Diagnosis and Treatment Terminology - Syndrome Part), with examples provided.

(4) Diagnosis and Treatment Method Information Table (): Records patient main symptoms, prescription names, usage methods, and lifestyle recommendations from physicians, referencing GB/T16751.3-1997 (TCM Clinical Diagnosis and Treatment Terminology - Treatment Methods Part).

(5) Imaging Examination Information Table (): Records imaging examination details when patients provide previous imaging data or when physicians require such information for diagnosis.

(6) Case Circulation Information Table (): Records case entry and export, clarifying responsible persons and transfer paths.

(7) Case Management Information Subset (): Records objective management of cases, including objective evaluation of case content.

The above tables organize the TCM diagnosis and treatment dataset, conforming to national and industry standards as much as possible while ensuring practical utility. As TCM continuously develops with new concepts emerging and old

ones being eliminated, these standards will be updated according to practical application.

3.2.2 Dataset Logical Structure Design Database structure refers to the structure of reasonably storing interrelated, logically structured data collections on computer storage devices. A database structure has multiple levels such as tables and fields. Due to the multi-layered and interdependent nature of TCM medical case information, statistical analysis must consider both overall perspective and key information for each part. As shown in [Figure 3: see original paper], this database structure design uses "case number" as the primary key to connect detailed information from both physicians and patients during diagnosis and treatment, as well as post-treatment preservation management information. This design avoids information overload in a single table, lengthy fields, and difficulty in locating information, enabling data analysts to clearly understand dataset structure and facilitating data access and processing.

3.3 Dataset Collection System Design This study uses C# application development for TCM diagnosis and treatment data collection system design and implementation, creating Windows form components and utilizing DataAdapter objects to enable display and interactive updates of database content. The DataAdapter object serves as a bridge between data sources and DataSet, both entering data into DataSet tables and returning DataSet changes to the data source. This enables the collection system to include not only data collection but also storage and modification functions, while data analysis can be performed within the database.

For field TCM diagnosis and treatment data collection, dataset entry interface and functional design are crucial, requiring high usability. A friendly information entry page must be built with logically arranged fields, enabling entry without frequent table switching. Entry boxes should include common automatic selection options to save time and provide convenience for physicians, patients, and data import personnel. While ensuring entry efficiency, some non-medical fields can be completed by patients themselves, with multi-terminal data synchronization designed to improve consultation efficiency. Specific fields for the four diagnostic methods are listed in sequence, allowing physicians to directly enter data in corresponding positions. When entering prescriptions and medical orders, drug names and usage requirements can be pre-entered into the database, allowing physicians to select defaults or modify usage requirements based on actual conditions without repetitive input. The data collection system interface is shown in [Figure 4: see original paper].

4 Empirical Research

The advantage of datasets lies in efficient data access and operation, with functions for rapid positioning, retrieval, and data interaction. To evaluate the rationality and flexibility of dataset construction, data was collected from both

existing online TCM cases and field diagnosis and treatment to initially form a dataset for empirical research.

4.1 Online Medical Case Data Collection

4.1.1 Online Medical Case Data Screening During online searches for TCM cases, it was found that numerous medical websites currently exist, but most TCM-related data are scattered and difficult to crawl. Different websites have vastly different data styles and fields for recording TCM cases and medical records. Considering data quality, the following screening principles were established for online case data:

(1) Cases lacking basic patient information are excluded. Cases missing necessary information such as patient gender and age are not included, as age and gender have important reference value for TCM diagnosis and subsequent analysis.

(2) Cases lacking main symptom descriptions are excluded. Since there are no unified standards or fixed formats for symptom description, each website and disease has different expression methods. To avoid overly strict requirements preventing data upload, websites often combine current medical history under "chief complaint" or "main symptoms." During trial entry, this portion of data was primarily collected into the patient basic information subset and diagnosis and treatment process information subset. Cases with vague, overly general, or questionable symptom descriptions were removed to avoid affecting subsequent data processing.

(3) Cases lacking treatment principles are excluded. The focus of every complete case and classical prescription is the treatment method and principles, which summarize the case and have the most statistical research significance. Some websites collect extensive information from medical texts and ancient books to explain difficult TCM terminology for public education, but omit or vaguely describe final treatment methods. Such symptom data are excluded.

(4) Non-case data is removed. Many medical consultation websites provide rich content to help online patients as much as possible, recording numerous fields unrelated to TCM case records, such as "whether this disease is covered by insurance," "best time for consultation," "preparation before consultation," and "treatment costs in tertiary hospitals." While helpful for patients, such information is irrelevant to case records and should be removed.

(5) Cases with intellectual property restrictions are excluded. Some websites have publicly shared information, some require notification for reprinting, and others explicitly prohibit copying. When entering online data, this information should be understood in advance, respecting personal privacy and intellectual property rights. Cases with intellectual property restric-

tions are excluded.

4.1.2 Trial Entry and Evaluation The main purpose of trial online data entry is to test dataset rationality and require standardized data entry. Trial data were selected from two open-source medical websites: "iiyi.com - Selected TCM Cases" and "TCM World - Medical Case Experience." These websites provide relatively comprehensive case information with fields similar to this study. The former contributed 54 manually collected selected TCM cases, while the latter contributed 47 cases after irrelevant information removal, totaling 101 cases saved in Excel format.

During trial entry, it was found that although online case data are abundant in quantity, many have questionable data (pointed out by website experts), vague information expression, or excessive data omission, resulting in low data utilization. Each website uses different fields for recording cases with varying emphases, essentially creating a "one field to multiple information tables" situation that causes great difficulty in data cleaning and prevents batch processing. However, this trial confirmed that data from existing online cases have matching fields in this dataset, with no phenomenon of relevant data having nowhere to be stored. This proves that the dataset design can meet the needs of collecting most case records from the internet.

4.2 Field TCM Diagnosis and Treatment Data Collection

Collecting data from field diagnosis and treatment better demonstrates this dataset's advantages compared to online data import. Due to TCM consultation characteristics, large amounts of implicit data cannot be collected during field diagnosis and treatment, such as "observing mental state," "tongue diagnosis," and "pulse diagnosis." TCM physicians often collect patient information related to conditions from subtle details but do not record them in cases due to efficiency concerns, leading to data loss.

4.2.1 Field Collection Methods Before field data collection, required research data should be evaluated in advance, planning data quantity and outpatient department information. Multiple methods can be employed to complete objectives:

(1) Physician entry: Physicians enter data in the collection interface during diagnosis and treatment, recording as much metadata-related information as possible for specific conditions. This method best ensures data quality but affects consultation efficiency and poses challenges for physicians with many patients.

(2) Physician assistant entry: Experienced TCM physicians typically have assistants for consultation management. Assistants can complete entry work while physicians verbally report information obtained through the

four diagnostic methods. Data review and organization can proceed simultaneously, making more efficient use of consultation time.

(3) Researcher entry: During dataset creation, project researchers can complete entry. One method involves on-site audio recording without disturbing physician consultation, with subsequent data entry based on recordings. Another involves researchers entering data on-site based on consultation progress. Researcher entry does not affect physician efficiency and researchers are proficient with the entry system, but audio recording requires prior communication about privacy protection, and entered data requires review by physicians or assistants.

(4) Multi-method integrated collection: Field collection offers high flexibility, allowing multiple methods to be employed simultaneously at consultation sites with different personnel collaborating from various angles to achieve maximum data accuracy and minimize omissions. However, collection processes require advance simulation and rehearsal, with attention to coordination that does not affect consultation efficiency.

4.2.2 Field Data Collection TCM diagnosis and treatment data collection cannot be accomplished in one day. It must neither affect normal consultation efficiency nor compromise dataset volume, requiring long-term accumulation to gradually improve and enrich the dataset. Before field data collection, researchers conducted simulated consultations to improve entry proficiency. During field collection, integrated multi-method collection was employed with collaboration between researchers, physicians, and assistants. Researchers entered patient basic information and final prescriptions; physicians verbally reported metadata-related information during diagnosis; assistants entered four diagnostic data and calibrated professional medical terminology. On-site audio recording was conducted without revealing personal privacy to ensure comprehensive and accurate data verification. Compared to web crawling, field-collected data guaranteed quality.

4.3 Dataset Structured Data Marking

Dataset construction cannot be done in isolation. An excellent dataset requires continuous testing and improvement through practical application. Nowadays, numerous datasets exist across various domains, creating difficulties in retrieval and search. Consequently, professional scientific dataset retrieval platforms have emerged, including comprehensive platforms like DCI and DatasetSearch, and biomedical-focused platforms like DataMed. To better evaluate research results, more dataset repositories are adopting schema.org and similar standards to describe datasets, with the types and coverage of datasets found through dataset search continuing to increase, facilitating data sharing and co-construction.

Considering subsequent dataset applications, this study adopts Google's dataset discovery method by adding schema.org standards to dataset description web-

pages and using structured testing tools to verify that structured markup has been added to dataset webpages. The uploaded dataset is shown in [Figure 5: see original paper]. In subsequent data collection, in addition to our research team, collaborators will be invited to jointly improve the dataset, further expanding data volume.

4.4 Data Application Prospects

Mi Hongying et al. summarized several analysis methods for TCM cases at the 14th International Collateral Disease Conference. Beyond personal comprehension analysis, statistical methods can be applied to process data within case collections, offering systematic and scientific advantages over personal comprehension. Zhang Xiaohang et al. explored applications of machine learning and deep learning methods in TCM diagnosis and treatment. Traditional machine learning algorithms such as clustering, classification, regression analysis, and association rules have been widely applied in TCM, while deep learning algorithms are closer to TCM's core and represent the major direction for future TCM development. The success of machine learning algorithms depends on large amounts of high-quality data. This dataset construction focuses on recording and analyzing each patient's individual conditions, addressing the limitation of existing datasets that focus only on specific diseases or ignore patient constitution, providing important support for data analysis in the TCM field.

Scientific data analysis is the main direction for medical and health development, but TCM faces issues of data loss and insufficient precision. This study constructs a TCM diagnosis and treatment dataset based on metadata to record the TCM four diagnostic methods and syndrome differentiation analysis in detail, forming a structured data system that breaks information silos. By involving patients, physicians, and researchers in data recording and organization, it promotes preservation, dissemination, and reuse of TCM diagnosis and treatment knowledge, which can then be analyzed and mined at deeper levels through database application technology to achieve knowledge sharing and inheritance. However, this study still has limitations: (1) Natural language processing technology should be applied to enhance batch processing capabilities for ancient and modern master TCM cases; (2) For complex case records, the time dimension should be better reflected to record patients' multi-year conditions in detail while minimizing information redundancy. Future research will build upon this dataset, accumulating experience through continuous empirical research, optimizing collection protocols, and improving data management systems to provide more comprehensive support for TCM diagnosis and treatment knowledge organization.

References

[1] Wang Dandan. Data quality control in the process of scientific data publication[J]. Library and Information Service, 2015, 59(23): 124-129.

[2] Lu Zhaolin. Complete Collection of Famous TCM Medical Cases Through Chinese History[M]. Beijing: Beijing Science and Technology Press, 2015.

[3] Wang Youhua, Lu Jingen, Liu Tao, et al. Knowledge discovery research in TCM medical cases[J]. Journal of Chinese Integrative Medicine, 2007, 5(4): 368-372.

[4] Liu Danhong, Zhang Lin, Yang Yue, et al. Overview of medical language and clinical data standardization[J]. Chinese Journal of Health Informatics and Management, 2014, 11(1): 14-17.

[5] Zhai Lincheng, Su Huaguan, Yang Weifeng, et al. Discussion on establishing and applying a quality control knowledge base for medical record front page data[J]. Journal of Medical Informatics, 2019, 40(5): 72-76.

[6] Xu Kun. Research on ontology-based scientific data curation platform[D]. Changchun: Jilin University, 2014.

[7] Jia Lirong, Nie Ying, Li Haiyan. Comparative study of DC, CKRM, and TCMLM[J]. Chinese Journal of Medical Library and Information Science, 2018, 27(12): 36-42.

[8] Zhao Yang, Cui Meng. Design principles and practical considerations for TCM literature metadata[J]. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2015, 17(10): 1978-1981.

[9] Liu Baojie, Chen Jin, Ma Lu. Practice in metadata structure design for TCM medical record database[J]. Journal of Medical Informatics, 2013, 34(12): 70-73.

[10] Sun Jing, Yang Fan, Deng Wenping, et al. Construction of an ontology-based TCM symptom knowledge representation model[J]. Journal of Medical Informatics, 2017, 38(2): 52-56.

[11] Liu Lihua, Hu Kai, Jin Shuigao. Research on metadata specifications for health information datasets[J]. Chinese Journal of Health Statistics, 2008(4): 363-365, 372.

[12] Liu Minjuan, Zhang Xuefu, Yan Yun, et al. Domain analysis dataset construction based on journal topic similarity: Methods and empirical study[J]. Library and Information Service, 2016, 60(10): 115-122.

[13] Shan Lianhui, Li Yong, Li Haicun, et al. Research on constructing medical field literature evaluation dataset based on DOI[J]. Journal of Medical Informatics, 2013, 34(2): 35-39, 44.

[14] Qiu Ruijin, Li Min, Hu Jiayuan, et al. Exploration of construction methods for core clinical safety evaluation datasets of post-marketing Chinese patent medicines[J]. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2018, 20(10): 1723-1728.

[15] Li Jiao, Guo Haihong, Guo Minjiang, et al. Quantitative study on thematic distribution and openness degree of open health and medical data from US and UK governments[J]. Library and Information Service, 2015, 59(20): 132-137.

[16] Ding E, Albuquerque D, Winter M, et al. Abstract 15983: A novel method of atrial fibrillation case identification and burden estimation using data in the electronic health record: data from the MIMIC-III dataset[J]. Circulation, 2018, 138(S1): A15983.

[17] Zeng Lei. Metadata and professional markup languages for knowledge representation in digital libraries[J]. Library and Information Service, 2002, 46(10): 14-22.

[18] Campbell A J, Fabrizio B, W L H, et al. Speaking the same language: using standardized terminology[J]. Journal of lower genital tract disease, 2016, 20(1): 8-10.

[19] Li Lu. Research on construction of medical case metadata model for TCM diagnosis and treatment knowledge base[D]. Changchun: Jilin University, 2020.

[20] Feng Lei. ISO releases first two international standards for TCM information[N]. China Traditional Chinese Medicine News, 2014-08-27(1).

[21] Jia Lirong, Liu Jing, Liu Lihong, et al. Research on disease-syndrome classification system construction for TCM clinical terminology system v2.0[J]. Chinese Journal of Medical Library and Information Science, 2018, 42(5): 8-12.

[22] Mi Hongying. Analysis methods for TCM medical cases[C]//Chinese Academy of Engineering, Chinese Association of Traditional Chinese Medicine, World Federation of Chinese Medicine Societies, et al. Proceedings of the 14th International Collateral Disease Conference. Jinan: Collateral Disease Branch of Chinese Association of Traditional Chinese Medicine, 2018: 240-244.

[23] Zhang Xiaohang, Shi Qinglei, Wang Bin, et al. Review of machine learning algorithms in TCM diagnosis and treatment[J]. Computer Science, 2018, 45(S2): 32-36.

Author Contributions

Ma Jie: Topic selection and formulation, research framework design, paper revision and finalization; Wang Jue: Dataset structure design, paper writing and revision; Sun Hengyu: Review and revision of TCM domain 专业知识; Li Lu: Investigation and formulation of metadata.

DataSet Construction of TCM Diagnosis and Treatment Based on Medical Case Metadata: Methods and Empirical Evidence

Ma Jie[1],[2], **Wang Jue**[1], **Sun Hengyu**[3], **Li Lu**[1] [1]School of Management, Jilin University, Changchun 130022 [2]Information Resources Research Center of Jilin

University, Changchun 130022 [3]The First Clinical Hospital of Jilin Academy of Traditional Chinese Medicine, Changchun 130022

Abstract: [Purpose/significance] TCM medical case is an important carrier to record the diagnosis and treatment data. It is of great significance for TCM diagnosis and treatment knowledge sharing, analysis and inheritance to collect medical case data based on metadata and form a dataset. [Method/process] We selected and determined TCM medical case metadata, and referred to the relational database for the conceptual and logical design of the dataset, proposed a method to collect, organize and store TCM medical case data to form a dataset, and conducted empirical research. [Result/conclusion] Network and field diagnosis and treatment data collection were conducted based on relatively complete TCM medical case metadata. The formed TCM diagnosis and treatment dataset can not only support multi-dimensional diagnosis and treatment information storage, sharing and retrieval, but also provide data sources for further diagnosis and treatment data mining.

Keywords: medical record; TCM medical record metadata; TCM diagnosis and treatment dataset

Note: Figure translations are in progress. See original paper for figures.