Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support

Xuezhong Zhou, Shibo Chen, Baoyan Liu, Runsun Zhang, Yinghui Wang, Ping Li, Yufeng Guo, Hua Zhang, Zhuye Gao, Xiufeng Yan

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
TCM Institute of Basic Clinic Medicine, China Academy of Chinese Medicine Sciences, Beijing 100700, China
China Academy of Chinese Medicine Sciences, Beijing 100700, China
Guangannmen Hospital, China Academy of Chinese Medicine Sciences, Beijing 100053, China
Beijing University of Chinese Medicine, Beijing 100029, China

1. Introduction

Traditional Chinese medicine (TCM) has a long history and the distinguished clinical effects. Different from the modern biomedical science, TCM has no general experimental practice in laboratory. In contrast, clinical practice or clinical experiment is the fundamental research activity of TCM. Hence, the new Chinese medical formulae and theoretical knowledge are not from laboratory but directly from the distilling of the daily clinical practice. Recently, TCM has been increasingly adopted as the complementary medical therapies for various kinds of diseases such as cancer [1], rheumatoid arthritis [2], leukemia [3] and migraine [4]. However, establishing a practical and rational efficacy assessment system is a vital issue for TCM to be widely accepted and used [5]. Identifying the common and repeatable medical knowledge or effective clinical strategies (e.g. herb prescriptions for a specific disease) from the clinical data is the

ARTICLE INFO

Article history:
Received 16 August 2008
Received in revised form 22 July 2009
Accepted 23 July 2009

Keywords:
Clinical data warehouse
Traditional Chinese medicine
Clinical data mining
Clinical decision support

ABSTRACT

Objective: Traditional Chinese medicine (TCM) is a scientific discipline, which develops the related theories from the long-term clinical practices. The large-scale clinical data are the core empirical knowledge source for TCM research. This paper introduces a clinical data warehouse (CDW) system, which incorporates the structured electronic medical record (SEMR) data for medical knowledge discovery and TCM clinical decision support (CDS).

Materials and methods: We have developed the clinical reference information model (RIM) and physical data model to manage the various information entities and their relationships in TCM clinical data. An extraction-transformation-loading (ETL) tool is implemented to integrate and normalize the clinical data from different operational data sources. The CDW includes online analytical processing (OLAP) and complex network analysis (CNA) components to explore the various clinical relationships. Furthermore, the data mining and CNA methods are used to discover the valuable clinical knowledge from the data.

Results: The CDW has integrated 20,000 TCM inpatient data and 20,000 outpatient data, which contains manifestations (e.g. symptoms, physical examinations and laboratory test results), diagnoses and prescriptions as the main information components. We propose a practical solution to accomplish the large-scale clinical data integration and preprocessing tasks. Meanwhile, we have developed over 400 OLAP reports to enable the multidimensional analysis of clinical data and the case-based CDS. We have successfully conducted several interesting data mining applications. Particularly, we use various classification methods, namely support vector machine, decision tree and Bayesian network, to discover the knowledge of syndrome differentiation. Furthermore, we have applied association rule and CNA to extract the useful acupuncture point and herb combination patterns from the clinical prescriptions.

Conclusion: A CDW system consisting of TCM clinical RIM, ETL, OLAP and data mining as the core components has been developed to facilitate the tasks of TCM knowledge discovery and CDS. We have conducted several OLAP and data mining tasks to explore the empirical knowledge from the TCM clinical data. The CDW platform would be a promising infrastructure to make full use of the TCM clinical data for scientific hypothesis generation, and promote the development of TCM from individualized empirical knowledge to large-scale evidence-based medicine.

© 2010 Elsevier B.V. All rights reserved.
essential solution to make TCM into the tested and regulated remedies [6].

In the post-genomic era, the large-scale biomedical information sources have been the key scientific data for modern biomedical research [7]. Likewise, construction of a large-scale clinical data set as biomedical information sources for modern biology is a fundamental scientific task in TCM research. The clinical practice with treatment based on syndrome differentiation (TBSD) is the basis of TCM clinical assessment and clinical study [8]. The huge clinical data storage is the firsthand and effective evidence for TCM clinical research. The clinical data record the tested TCM data in clinical practice and come directly from the patient treatment. Hence, it has become a significant and vital research issue to develop an intelligent analysis platform to manage the large-scale TCM clinical data, and help efficient discovery of the hidden knowledge from those heterogeneous information sources. In general, it is significant to build the TCM clinical wet–dry mode [9] that fully uses the clinical data to promote the TCM studies. Because the information content that is considered in the TCM clinical practice is huge and high dimensional (includes not only the traditional disease conditions like symptom and sign, but also the environmental and social factors), induction of the stable and common relationships between manifestations, diagnoses and treatments are very difficult. Hence, the knowledge discovery task to generate tractable hypothesis should be emphasized before a proper random controlled clinical trial is performed.

Data warehouse [10] is a technical solution for immense data storage, management and processing. The increased demands on financial analysis [11], disease control [12], clinical decision process [13], adverse drug events control [14], laboratory test data analysis [15], information feedback for hospital practice management [16], large-scale radiologic and pathologic data management [17] and clinical data mining (DM) [18] in healthcare have given rise to the research and development of clinical data warehouse (CDW). Clinical data warehousing is a difficult systematic task with many particular complicated issues, such as many-to-many relationships, entity-attribute-value (EAV) data structure and bi-temporal data [19]. Data integration of medical data storage is challenging, hence the data warehouse architectures [20] were studied to propose practical solutions to tackle data integration issues. Based on the TCM clinical reference information model (RIM), we introduce a data warehouse solution to process and analyze the large-scale TCM clinical data sources.

Compared with the clinical data of modern medicine, TCM clinical data have the distinct and significant information contents, such as symptom and sign, syndrome, formula and herb, as the core components. Moreover, the symptom and sign information with systematic description is the foundational information for syndrome diagnosis. Therefore, the medical record containing symptom and sign should be structured and relationally stored. However, the structured data entry of electronic medical record (EMR) is still a research issue that needs to be further explored [21], and the manual structured medical record data entry is a labor-intensive task. Hence, although the free-text EMR data need further information extraction are in huge scale, the collected structured EMR (SEMR) data in medical domain are still very rare.

In 2002, we have developed a TCM SEMR system [22], which stores the SEMR data (e.g. the clinical events and entities contained in the chief complaint, histories and progress notes) in relational database. Using the SEMR system, we have manually collected about 20,000 inpatients data of diabetes, coronary heart disease (CHD) and stroke from the TCM hospitals (a kind of hospital in China that has TCM as the main clinical approach) or TCM wards. In addition, over 20,000 outpatient data cases, which record the outpatient clinical encounters of 20 high-experienced TCM physicians in Beijing, are collected. Furthermore, we have a systematic study on the TCM clinical terminology and nomenclature [23] to facilitate the data entry and standardized representation of structured information elements. To utilize and analyze the TCM clinical data for research purposes, we have developed a CDW system to support the medical knowledge discovery and the clinical decision-making. By comprehensive analysis of the characteristics of TCM clinical data structure and the subjects of TCM clinical research topics, we have designed the RIM, the physical data model and the multidimensional data model for CDW. Meanwhile, we have developed an extraction-transformation-loading (ETL) tool, called medical integrator (MI), to take the tasks of clinical data integration, data cleaning and preprocessing. We have also integrated the DM systems, namely Oracle data miner (ODMiner) [24] and Weka [25], and the business intelligence system (BusinessObjects) [26] to implement a TCM clinical intelligence analysis platform with data mining and online analytical processing (OLAP) abilities. Moreover, we have studied on the scale-free network phenomena of TCM clinical data [27], and found that it will be a promising approach to analyze the TCM clinical data from complex network perspective. Currently, we have developed a complex network analysis (CNA) system, called TCMNetBench, to automatically construct the corresponding network models and directly analyze the data from the CDW.

The rest of this paper is arranged as follows. The core components of the TCM CDW, such as the infrastructure, RIM, ETL and data analysis component, are introduced in Section 2. Due to the various expressions of terms in daily clinical practice, it is a nontrivial task to utilize the SEMR data for knowledge discovery and decision support. The data integration and preprocessing are necessary steps to assure the data quality, which is discussed in Section 3. Then we focus on describing the OLAP multidimensional analysis that supports the clinical decision-making in Section 4. We discuss the DM issues in TCM clinical data and introduce the pilot TCM clinical DM applications such as classification methods for syndrome differentiation (e.g. the differentiation of qi stagnation syndrome), association rule mining and CNA for acupuncture point and herb combination knowledge in the clinical prescriptions, in Section 5. Finally, we conclude and discuss the future work in Section 6.

2. TCM clinical data warehouse

There are several CDW systems developed to utilize the potential knowledge in clinical data. However, most of the previous work focused on the medical information management [12,16], and the processed EMR data are in free text [28]. The symptom and sign information in the EMR was not considered. We aim to build a CDW for TCM clinical and theoretical research purpose, which takes the SEMR data as the core data source. The clinical information, such as chief complaint, is structured, and the terminological entities and the related descriptions (e.g. the symptoms, such as thirstiness, headache, poor appetite and chest distress) in the sentences are extracted and stored as EAV schema.

2.1. The infrastructure of TCM clinical data warehouse

The TCM CDW system is designed based on Java and J2EE platforms. The technology infrastructure of TCM CDW is depicted in Fig. 1. It shows that the infrastructure aims to integrate different operational data sources (e.g. SQL Server, Oracle, DB2) by using a self-developed ETL tool. Due to the heterogeneous operational data sources, we use a series of metadata information tables to record the metadata (e.g. database type, hospital information, physician information, data content description and transforming information) of the different data sources.
The data storage management is supported by Oracle (currently, we use Oracle 10G as the database server), also the analysis and query service is mainly supported by Business Objects (BO). BO has the design and analysis clients like Crystal Report, Web and Desktop Intelligence, and Dashboard and Performance manager to implement the OLAP functions. Meanwhile, we integrate ODMiner and the Weka machine learning platform to perform the online DM tasks. The TCM network workbench is a CNA system to visualize and analyze the TCM clinical data in network model. Therefore, the infrastructure builds a technological framework for huge TCM clinical data integration, preprocessing, management, online analysis and DM. As a platform for TCM clinical research, TCM CDW system also can directly provide prepared data for the statistics software (e.g. SPSS, SAS and STATISTICA) to make possible statistical analysis and test. Hence, from the application perspective, TCM CDW proposes an integrative functional platform with support for raw clinical data integration and data preprocessing, OLAP, DM and statistics analysis tasks (Fig. 2).

2.2. TCM clinical reference information model

The information model design is the vital step of TCM CDW development. Medical information model like HL7 RIM [29] is a very complicated system with various classes and relationships. The objectives of HL7 RIM are to support the medical operational processes, particularly support the information exchanging between different medical information systems. The semantic network of unified medical language system (UMLS) [30] is considered as the distinguished medical ontology in modern biomedical science. The semantic types and structures proposed a global conceptual view of the medical terminologies. The main aim of UMLS is to bridge the gaps between different terminological systems used in the medical literature.

However, the RIM of TCM CDW focuses on the information content that would be analyzed and used in the clinical and theoretical research. Hence, the classification and definition of the information generated by the TCM clinical processes, is the emphasis of our work. We consider TCM clinical process as a dynamic system with two core entities, namely physician and patient, and three core information elements, namely symptom (it generally includes all the clinical manifestations of patient such as physical examination and laboratory test), disease/syndrome and treatment. The symptom information element is regarded as a relatively objective disease phenomenon, whereas, disease/syndrome is one type of human morbid state, which is identified by a physician. Meanwhile, the TCM treatment is a clinical event that aims to make patient healthful. Actually, time structure and temporal reasoning is an essential feature in medicine [31–32]. Therefore, we take the abstraction of these five core information elements in the temporal conditions and design the global conceptual framework of TCM RIM (Fig. 3) for CDW.

We consider that the main objective of TCM clinical research is to study on the relationships between different entities in each event and the relationships between different events. Therefore, we regard the clinical information as various kinds of events (phenomenon and activity), and in each event there have several conceptual entities and physical entities participated at a specific time point or in a time interval.

The event class defines the clinical phenomena and activities that happen in the temporal conditions. Each sub-classes of the event would have the time as the essential attribute. The phenomenon class defines both the consequential clinical findings that are directly discovered by clinical observation or measurement, and the clinical states that are recognized by physicians. The information of phenomenon class is always generated by the corresponding activity class in the event class. However, we emphasize defining the phenomenon class to represent the complicated information of patient manifestations in the sections of EMR, such as the chief complaint, histories, physical examination and laboratory test. For example, the symptom and sign descriptions in the histories of present illness, such as “the patient has recurrent chest discomfort and palpitation for 2 months, and occasionally has dizziness in the morning”, often contain various kinds of significant information to be considered in the diagnosis and treatment. Because of the variable and various kinds of manifestation information in clinical data, the EAV data model is preferred in the health care information system. We also have implemented the TCM SEMR system based on EAV data model. In
the EAV data model, each row of a table corresponds to an EAV triple: an entity, an attribute, and the attribute value. For example, the entity “patient” can have the attribute “postprandial blood sugar”, a laboratory test result with a float value of 13.5. The EAV model facilitates the regular operations in health care information system. In addition, it is flexible to be used in the data analysis applications. Hence, we also use EAV model to materialize all the phenomenon classes in physical data tables.

The activity class is defined to capture the information of clinical process and treatment. It is similar with the act class in HL7 RIM [29] except that the activity class focuses on defining the factual actions, which have happened, it does not include the actions of happening or scheduled to happen. This is also fit for the research purpose of the clinical RIM since we aim to propose a general RIM for data analysis and clinical research. The information of the activity class will record the information content like “when and where the activity is performed”, “who performs the activity” and “what has been conducted in the activity”. The activity class includes the sub-classes such as registration, examination, laboratory testing, diagnosis, treatment and follow-up visit. Some of the result information of the sub-classes, namely examination, laboratory testing, diagnosis, treatment and follow-up visit will be recorded in the corresponding sub-classes of phenomenon. However, the prescriptions and therapies of treatment activity are directly defined as the sub-classes of treatment class because these information categories are regarded as the concrete contents of treatment.

There are two kinds of entity classes, namely conceptual entity and physical entity, in the clinical RIM. These two entity classes aim to define the elementary information classes, such as physician, patient, symptom, and disease in the clinical data. The conceptual entity class defines the conceptual or functional abstract entities, and the physical entity class represents the substantial and physical objects in the clinical data. The conceptual entity class gets the standardized terminology support from the existed clinical terminology systems such as UTCMLS [33], TCM clinical terminology and nomenclature [23] and ICD-10 (it is inherited from the SEMR system). Because of the mixture of TCM and modern medical concepts and methods in current TCM clinical processes, some sub-classes of conceptual entity class like disease are also the mixture of TCM and modern medical classes. For example, we have defined two distinct disease classes, namely TCM disease and modern disease, in the model. The former represents the disease concept in TCM, while the latter is the modern medical concept. It is worth mentioning that the conceptual entity classes would be materialized as dictionary tables in the physical data model in data warehouse. The physical entity class defines the physical and substantial objects involved in the clinical practices. The sub-classes like patient, physician and medication are the core entities. The medication class includes the herb, Chinese traditional patent medicine and drug sub-classes. These sub-classes would be materialized to dictionary tables, in which some of the fields get the standardized terminology support from the conceptual entity class, such as herb effect and herb nature. Therefore, we construct the TCM RIM with an abstract level and granularity fitting for the TCM theoretical research purpose. The information classes like roles of HL7 RIM, which are mainly to support the clinical operation management, are not considered. We have the more detailed description of the RIM in the work [27]. Based on the defined information model, we have designed the physical data model to manage the detailed TCM clinical data from the heterogeneous operational data sources. The physical data model has 18 core physical data tables such as patient, clinical registration, diagnosis, clinical finding, laboratory test, order,
clinical formula, drug, herb, progress notes and scale. These core tables provide the storage schema for the detailed clinical data. Furthermore, to support the multidimensional analysis tasks such as OLAP, we have designed several core relational multidimensional data models as the basis of data marts. There are several significant subject analysis applications for TCM clinical research purpose, which have the corresponding relational multidimensional data models such as clinical formula, clinical diagnosis and clinical finding. For example, the clinical formula relational multidimensional data model is defined in snowflake schema, which has a clinical formula fact table with several related dimensional tables like patient, physician, time, diagnosis, herb and therapeutic method. The herb dimension is further normalized into multiple related tables such as herb nature and flavor table, herb efficacy table and herb channel entry table. The formula data model is used for the clinical formula prescription data analysis. The frequent formulae, herbs (also the herb properties) and therapeutic methods for a specific disease could be explored based on the clinical formula data model. In Sections 4 and 5, we will discuss the related OLAP analysis and DM applications in detail. The practical results show that the RIM and multidimensional data model can support well for the clinical analysis applications.

2.3. Medical integrator

The ETL is the core component of a successful data warehouse system. Due to the requirement of complex clinical data structure transformation, flexible data checking, heterogeneous data sources integration and terminological standardization processing, even the commercial ETL systems can not fit well for the tasks. Using Java and Eclipse standard widget toolkit, we develop the ETL tool, MI, to implement the required functions. It has the key components such as data connection configuration, data checking, data integration (e.g. operational data source consolidation, data
transformation and loading), data cleaning, data standardization and data analysis interface. Due to the distributed SEMR data collecting in different hospitals and wards, the data integration component has been developed to integrate the multiple operational data sources (e.g. inpatient and outpatient SEMR data) to one unified data structure, and load the transformed data into the TCM CDW. Because the EAV data structure is used in the SEMR system and strict data entry control is not applicable in the clinical practice, we should check each value of the original data to avoid loading the invalid data into the data warehouse. Furthermore, MI has focused on the particular functions like data standardization and data analysis interface. Data standardization process mainly concerns the standardization of the terminological data like clinical finding, diagnosis and treatment (e.g. herb name and description phrase of the therapeutic methods). The data analysis interface implements the functions of preparing the data for OLAP analysis and DM. It includes the components of data transformation of detailed physical data to multidimensional data, the transformation of EAV data to conventional data and the data exporting for statistical software. MI performs significant tasks to integrate and preprocess the heterogeneous TCM clinical data. We will have a further discussion of the data integration and preprocessing issues in Section 3.

2.4. Data analysis components

Based on the multidimensional data model and ETL preprocessing, we have prepared the clinical data for the OLAP analysis and data mining tasks. We develop the OLAP analysis applications based on BO platform. To make the report-designing task available for clinical experts, we designed several semantic layers to map the physical data structures to domain knowledge categories by using BO Designer. Based on the semantic layers, even clinical experts can design their own reports according to their personal requirements. In addition, we have integrated the DM systems, such as ODMiner and Weka, to the CDW. The integrated two DM systems have the online data access ability to the clinical database. This has a great help to facilitate the performing of the DM tasks.

Furthermore, we have developed a CNA system, called TCMNetBench, for TCM clinical data. TCMNetBench filters the data set directly from the data warehouse, and automatically constructs several interesting TCM clinical networks, such as herb combination network, symptom co-occurrence network and complication disease network. The constructed networks could be displayed and analyzed with manually tuned parameters. We implement the network visualization and analysis functions using JUNG graph package (http://jung.sourceforge.net/index.html). Using TCMNetBench, the clinical researchers could filter the data set from the huge CDW by the expected combinational conditions, such as inpatient of type 2 diabetes with thirst symptom and the outpatient data of a specific TCM physician with Xiao Chaihu decoction (XCD) treatment. Then the objective networks would be automatically constructed and stored based on the filtered data set. The metadata of the constructed network is also recorded in a table. Finally, the stored network data could be explored and visualized by searching the metadata table. Several properties like node size, node shape, node color, edge style, layout and scale of the visualized network can be configured and changed. The core sub-networks could be filtered and clustered by the ranking methods (e.g. degree and betweenness) and community identification algorithms respectively. To address community discovery from the dense networks, we have implemented a novel algorithm to extract the core hierarchical sub-networks from the constructed large-scale networks [34].

3. Data integration and preprocessing

Multiple heterogeneous clinical data sources with diverse clinical data standards (e.g. ICD-10, SNOMED-CT) produce challenges for data integration. Three dimensions, namely intra-facility, cross-facility and time, are considered in clinical informatics to promote the reliability of the clinical integration [35]. TCM CDW contains both inpatient and outpatient SEMR data, whose operational databases have different physical data tables to store the structured data. Moreover, due to the flexible term descriptions and expressions in clinical practice of different hospitals and wards, novel term expressions would be added to the database in the individual SEMR systems. Therefore, we should construct a unified CDW that integrates the data from the heterogeneous data sources with multiple clinical data standards. The uniqueness of medical DM [36] proposes special requirements for clinical data preprocessing. The poor data quality, inconsistent representation, and the requirement of data structure transformation from online transactional process to data analysis and the complicated domain knowledge make data preprocessing of clinical DM a labor intensive and error prone task [37].

Due to the multiplex TCM theories (e.g. yin-yang, five phases, zang-fu viscera and channel theories) and the diverse clinical operations of TCM clinical work, it is important to address commonality with heterarchical levels and granularities. Furthermore, in TCM hospitals, it is popular to use the integrated traditional Chinese and western medicine in the inpatient encounters. This means that TCM clinical data would include both TCM and modern biomedical information. Therefore, the clinical data preprocessing should address the standardization and transformation of both of the two data sources. To have a reliable and flexible data preparation process, we conduct the data integration and preprocessing by two steps. The information pipeline of the data integration and preprocessing is depicted in Fig. 4.

In the first step, we would merge a specific kind of data sources (e.g. inpatient of diabetes, outpatient of a TCM physician) from different data collecting sites into one single database. Then the merged single database is integrated into the CDW by necessary data checking (e.g. data completeness, continuity and accuracy) and transformation. There are over 400 physical tables in the SEMR data source. We rearrange and extract the information from these tables to the 18 core detailed physical tables in CDW. Through the data mapping and transformation, we will keep the consistent data storage in the lowest information granularity in the detailed data warehouse. Hence, the detailed data warehouse becomes an integrated copy of the original data sources.

In the second step, we conduct the data transformation and preprocessing (e.g. data standardization, data conversion) to generate prepared data for OLAP and DM. The data transformation process conducts the transformation of detailed physical data to multidimensional data, and the EAV data to conventional coded data (each attribute becomes a data field and the related values are normalized to coded data). The data standardization process focuses on the standardization of the terminological data such as symptom, diagnosis and herb treatment with the clinical terminology system support. In TCM, it is notable that the information of symptom and sign plays a vital role in both diagnosis and treatment. For example, a CHD patient with chief manifestations of “intermittent stinging pain in the chest and aggravated at night, palpitation, restlessness, dark purple tongue, and deep and unsmooth pulse” would be differentiated as stagnation of heart blood syndrome. Thereafter, the modified Xiefu Zhiyu Tang (decoction for removing blood stasis in chest) would be prescribed for the treatment. Furthermore, due to the comprehensive approach used to acquire and describe the symptom and sign information
in TCM clinical data is complicated and keeps with multidimensional, multiple and heterogeneous characteristics. Therefore, it is important to propose an effective method to preprocess the symptom and sign data.

We use a rule-based batch processing approach to perform the data standardization tasks. About 8 rule tables are designed to store the different kinds of standardization rules. The TCM clinical experts should conduct the rule editing tasks and import the rules into the corresponding tables by using MI. To keep the original data for different analysis applications, we let MI build the necessary additional tables to store the processed data, and provide a standardized data set for different data analysis applications. For example, in the symptom standardization process, the domain experts need edit four kinds of transformation rules, which instruct the processes of noise data cleaning, unified term description, terminological granularity unification and synonymous unification. The result of symptom standardization is the terminological phrases with the unified concept.

The EAV structure [19,38] is the preferred choice in clinical data model. However, most statistical and DM systems require conventional flat style data and encoded data. Therefore, to prepare the clinical data to be seamlessly used by the statistical and DM systems, we have developed several key functions, such as automatic encode processing, EAV to flat schema conversion and data exporting, in MI system. By using the functions of MI system, it enables a good preparation of data set with high quality for various data analysis tasks.

4. Online analytical processing for clinical decision support

TCM clinical practice is an innovative and complicated procedure, in which TCM physicians should have a deduction of diagnosis and treatment based on their empirical knowledge and the basic TCM theories. It will be a great help if the established clinical data could provide informative clues or similar cases for TCM physicians when they are in the face of the difficult and intractable cases. Clinical decision support (CDS) provides clinicians, staff, patients, or other individuals with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care [39]. Based on the multidimensional data schema and BO semantic layers, we have developed 10 OLAP subject analysis applications with over 400 analysis reports to explore the clinical data and provide valuable information for CDS. The subjects mainly focus on two types of TCM clinical knowledge, namely the empirical diagnosis and treatment knowledge of the high-experienced TCM physicians, and the clinical features of vital chronic diseases like diabetes, stroke and CHD. The analysis reports, such as, the global data profile of a specific TCM physician, the frequency of clinical herb and formula, and the relationships among medical entities (e.g. clinical findings, syndromes, diseases and complications), have been developed and can be used by the authorized web users. Besides the interactive browsing of the reports, the user can also export the results as Excel or PDF file for further investigation.

As shown in Table 1, the report of the global data profile of a high-experienced TCM physician proposes the information about the total number of patients, encounter times, and the frequency of disease, syndrome, symptom, herb and formula, etc. The global data profile provides the baseline information of the clinical data related to a specific physician. It shows that there are 310 outpatient cases with 1135 encounters, and the clinical data are mainly on the TCM diseases such as thoracic obstruction of qi, gastric pain, palpitation and vertigo. While the modern diseases treated by the TCM physician are rather broad as it shows that there are about 151 different modern diseases such as gastritis, CHD and hypertension in the 1135 outpatient encounters. The other information like clinical herb frequency (in Table 1) could be drilled down to the detailed data (Table 2). It shows that there are 390 different herbs used by the TCM physician. The most frequent herb is Indian bread with 506 times and the average, maximum and minimum dosage is 18.2, 40 and 8 g, respectively. The global data profile not only describes a general view of the clinical data of a high-experienced TCM physician, but also proposes the useful information for young TCM physicians to learn the basic empirical knowledge, such as the clinical herb prescription knowledge, from a high-experienced TCM physician.
There are mainly two types of empirical knowledge in the TCM clinical data. One is about how to make TCM diagnosis, and the other is about how to make prescription. The OLAP portal provides the two types of knowledge in multiple dimensions and granularities. For example, in Fig. 5, it shows that the symptoms such as wiry and slow pulse, chest distress, shortness of breath, white tongue coating are the frequent symptoms of qi stagnation syndrome (QIS syndrome) with 75 outpatient instances of a high-experienced TCM physician. In addition, when the QIS syndrome is diagnosed, the rational herb prescription is vital for treatment. Fig. 6 shows the frequent herbs prescribed by the physician for the CHD patients with QIS syndrome. It suggests that the herbs like indian bread, dried tangerine peel, liquorice and largehead atractylodes rhizome, etc. would be prescribed in the corresponding formula therapies.

Many effective classical formulae have been developed during the long history of TCM practice. The classical formulae, such as XCD and Xiaoyao San, are widely used in the current clinical practice. However, how to apply the formulae with herb modifications to the appropriate disease conditions is a nontrivial clinical procedure. We can display this kind of knowledge by using the outpatient cases of the CDW. The related diseases treated by XCD (based on the data of a high-experienced TCM physician, who is proficient in the use of XCD) are displayed in Fig. 7. It shows that the most frequent diseases are chronic gastritis, upper respiratory tract infection (URI), hypotension, etc. This means that XCD with the herb modifications is frequently used for treating these diseases. However, TBSD is the main principle of TCM herb prescription for individualized patients. Hence, an effectual herb prescription is possible only if the TCM physician knows the specific syndromes of an individualized patient. Table 3 shows the related syndromes and the concomitant formulae while XCD is prescribed for the diseases (e.g. superficial gastritis, URI and Neurosis). It shows that the XCD with Zuojin Wan modifications would be prescribed for the chronic gastritis patients only if they have the disharmony between the liver and spleen (GPBT) syndrome. TCM proposes the individualized therapies for patient with the same disease. Hence, the formula-syndrome association knowledge of a specific disease will give the other TCM physicians a practical reference for the clinical treatment.
5. TCM clinical knowledge discovery

As a DM domain with high rewards and uniqueness [36], medical DM has been a hot research topic in recent years. Mining over medical, health or clinical data is considered as the most difficult domain for DM [40]. There is plenty of related research in clinical DM [18,41–43] of modern biomedicine. The study on the methods and data analysis of laboratory test results, medical images and demographics to make diagnosis or prognosis prediction are the main research topic [44–46]. For example, Prather et al. [47] have a preliminary study on the medical DM in CDW, and 3902 obstetrical patients were evaluated for factors potentially contributing to preterm birth using exploratory factor analysis early in 1997. Breaulta et al. [48] use the classification tree method to analyze the diabetic data warehouse with 30,383 patients. The novel association between bad glycemic control and younger age was discovered, which would be useful for clinicians and administrators. As clinical data contain multiple types of attributes (e.g. numerical attributes and categorical attributes), Hirano et al. [49] conducted the experiment on different similarity measures in clustering of medical data, and showed that Ward’s similarity measure would be a good method to get better clustering on the medical data sets. Mullinsa et al. [50] apply the three unsupervised learning methods of a new DM software, HealthMiner, to a large cohort of 667,000 inpatient and outpatient digital records from an academic medical system. The initial results show that it is feasible to combine and apply large-scale DM search tools to complex clinical data sets. Cho et al. [51] applied several machine learning techniques, such as support vector machine (SVM) classification and feature selection methods, to predict the onset of diabetic nephropathy using an irregular and unbalanced diabetes data set.

Compared with the clinical DM research in modern biomedicine, TCM clinical DM only becomes the hot topics in recent years. The related work of TCM knowledge discovery has been reviewed by Feng et al. [52] and Lukman et al. [53]. Recently, Zha et al. [54] used neural network to predict the role of diagnostic information in treatment efficacy of rheumatoid arthritis. Chen et al. [55] had a comparative study on the five classification methods (e.g. SVM, neural network and decision tree) for syndrome differentiation of CHD using 1069 clinical epidemiology survey cases. The result shows that SVM performs best in the prediction task. Tang et al.
have worked on the mining of the elder TCM masters’ knowledge to visualize their thoughts and facilitate the knowledge transfer. To illuminate the statistical foundation and objective diagnosis standards of syndrome differentiation, Zhang et al. proposed the latent tree model to learn the diagnosis structures from the TCM clinical data sets with symptom variables in an unsupervised way. The study showed that there exist natural clusters in the data sets, such as the data set of kidney deficiency syndrome with 2600 cases, which correspond well to the syndrome types. Hence, it is promising to uncover the significant knowledge from large-scale clinical data while appropriate DM methods and high quality data are used. Actually, inductive analysis of the empirical data from clinical practice is a key step for TCM clinical research. Moreover, study on the relationships between primary conceptual medical elements like disease, syndrome, symptom, herb and formula is the central topic of TCM clinical research. The DM would be an efficient approach to get the related knowledge and evidences from the practical clinical data. Corresponding to the TCM clinical research topics, we have a depiction of the clinical DM topics in Fig. 8. It shows that there are various types of knowledge and topics, such as syndrome differentiation, herb combination, formula and syndrome relationships, and syndrome epidemiology, could be explored by DM methods. Currently, the related research focuses on the study of syndrome differentiation (generally includes syndrome epidemiology) and herb combination. However, due to the practical integrated clinical practice in TCM hospitals, there are various interesting topics could be further explored and studied on the relationships among manifestation, diagnosis and clinical therapy. The TCM CDW provides a well-prepared information source for clinical DM research.

We have successfully conducted several preliminary TCM clinical data analysis studies like acupuncture prescription knowledge discovery [59], the relationship between formula (herbs) and syndrome of T2DM affiliated metabolic syndrome (DAMS) [60], herb treatment for T2DM [61], cluster analysis on the syndrome types in patients with acute myocardial infarction [62], and the GPBT syndrome differentiation [63]. All the previous work has got the clinical useful results.

We have conducted the classification study on QIS syndrome differentiation of a high-experienced TCM physician to validate the

Table 3
Part of the detailed adaptation diseases with syndrome differentiation of XCD in the clinical prescriptions of a TCM physician. The number in the parenthesis represents the frequency of the related entities (e.g. formula combination and syndrome combination).

<table>
<thead>
<tr>
<th>The frequency of modern diseases</th>
<th>The frequency of formula combinations and syndrome combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superficial gastritis (4)</td>
<td>Formula combination: (XCD, Zuojin Wan) (2). Syndrome combinations: Liver fire attacking the stomach and GPBT (1); Liver qi attacking the stomach and disharmony between liver and stomach (1). Formula combination: (XCD, Sini San, Zuojin Wan, Shaoyao Gancao decocction, Shenjiang San) (1) Syndrome combinations: Disharmony between liver and stomach (1). Formula combination: (XCD, Zuojin Wan, Xiao Xianxiong decoction) (1). Syndrome combinations: Liver qi attacking the stomach, dysfunction of the spleen in transporting (1).</td>
</tr>
<tr>
<td>Upper respiratory tract infection (3)</td>
<td>Formula combination: (XCD, Shenjiang San) (2). Syndrome combinations: Pathogen invading Shaoyang and accompanied with wind and fire (2). Formula combination: (XCD) (1). Syndrome combinations: Pathogen sticking Shaoyang, healthy qi not filling (1).</td>
</tr>
</tbody>
</table>

Fig. 8. The main topics of TCM clinical data mining. The red thick lines show the current TCM research topics. The green thick lines show the clinical data mining topics of modern biomedicine. The grey thin lines and yellow thin lines indicate the promising topics that should be further explored.
Table 4
The experimental results of classification methods. The chief symptoms are ordered by attribute weight, prediction weight in SVM and ADTree respectively. The chief symptoms of Bayesian network are ranked by the ratio of the probability distribution of symptom on QIS syndrome and non-QIS syndrome.

<table>
<thead>
<tr>
<th>Classification method</th>
<th>Test mode</th>
<th>Prediction accuracy</th>
<th>The five suggested chief symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>Evaluate on training data</td>
<td>0.864</td>
<td>Costal pain (1.7166), back pain (1.1961), chest pain and oppression (1.0361), breathlessness (0.7063) and gastric pain (0.6704)</td>
</tr>
<tr>
<td></td>
<td>10 cross-evaluation</td>
<td>0.812</td>
<td>Costal pain (1.7041), back pain (1.1745), chest pain and oppression (0.9662), breathlessness (0.7833) and gastric pain (0.6959)</td>
</tr>
<tr>
<td>ADTree</td>
<td>Evaluate on training data</td>
<td>0.841</td>
<td>Chest pain and oppression (0.732), gastric pain (0.592), dyspnea (0.402) and slow pulse (0.383)</td>
</tr>
<tr>
<td></td>
<td>10 cross-evaluation</td>
<td>0.814</td>
<td>Chest pain and oppression (0.732), gastric pain (0.592), dyspnea (0.402) and slow pulse (0.383)</td>
</tr>
<tr>
<td>Bayesian network</td>
<td>Evaluate on training data</td>
<td>0.845</td>
<td>Costal pain (0.042/0.006), back pain (0.030/0.006), chest pain and oppression (0.042/0.011), gastric pain (0.077/0.023) and breathlessness (0.053/0.021)</td>
</tr>
<tr>
<td></td>
<td>10 cross-evaluation</td>
<td>0.806</td>
<td>Costal pain (0.042/0.006), back pain (0.030/0.006), chest pain and oppression (0.042/0.011), gastric pain (0.077/0.023) and breathlessness (0.053/0.021)</td>
</tr>
</tbody>
</table>

The herbs in the same formula are organized according to the principle of formula theories such as the monarch, minister, assistant and guide theory, to make a prescription with systematic efficacy. Herb combination and the basic formula with common herb combinations play a key role in the efficient clinical prescriptions. Hence, discovery of the common herb combinations from large-scale clinical formulae has been a significant research topic in TCM DM. Frequent itemset and association rule have been used as the general DM methods to find the interesting herb combinations. We have conducted several studies using association rule mining method, such as acupuncture prescription knowledge discovery [59]. The work focuses on the empirical clinical acupuncture prescription of a doctor in acupuncture department of Guanganmen hospital, Beijing, China. Using the association rule mining method (Apriori) in Weka, we got 18 common acupuncture formulae from the 1697 clinical prescriptions. The doctor indicates that one of the 18 acupuncture formulae is not a fixed prescription in his clinical practice. Therefore, finally, we get 17 useful acupuncture formulae (with name, acupuncture point composition, modifications, main efficacy, etc.) for different disease conditions, which reflect the empirical knowledge of the doctor. However, association rule mining has the shortcoming that it often generates large amount of rules due to the limited number of herbs (not over 1000 herbs in clinical use) and acupuncture points (not over 400 points in the clinical use), and the almost unlimited clinical prescriptions. Furthermore, the rules generated by association rule method are often in low granularity, thus association rule method cannot capture the skeleton herb combination knowledge in the large-scale prescriptions. We have implemented the hierarchical core sub-network structure discovery algorithm [34] in the CNA workbench. The method builds the large-scale herb combination network from the complete connected sub-networks with the element herb as node of one formula, and then extracts the hierarchical core herb combinations from the constructed herb combination network. We have applied the method to several specific TCM prescription data sets extracted from the CDW, and got the clinical meaningful common herb combination patterns. We take the knowledge discovery from the two data sets, namely the herb prescriptions of DAMS and the herb prescriptions of GPBT syndrome, to demonstrate the interesting analysis results.

Metabolic syndrome\(^1\) is one of the significant complications of T2DM, which may have the symptoms and features of high blood pressure, central obesity, and fasting hyperglycemia. Distilling the common herb combination knowledge from the large numbers of clinical prescriptions for DAMS will help guide the efficient herb treatment. There are 188 inpatient cases of DAMS that have been treated with herb prescriptions, are filtered from the 5000 over diabetes inpatient data according to the WHO standard. The data set has totally 752 herb treatment prescribed in the different encounters. The herb prescriptions have 320 different herbs and the constructed weighted herb combination network has 9541 edges (two-item herb relationships). The extracted core herb combinations and main herb modifications of DAMS prescriptions are displayed in the right part of Fig. 9. The core herb combinations contain the herbs, such as Chinese angelica, dwarf lilyturf tuber, milkvetch root, Chinese magnoliavine fruit and figwort root. According to the efficacies of these five herbs, it shows that the organization principle of core herb prescription for DAMS is to replenish qi and nourish yin (one type of the TCM therapeutic methods). This reflects the basic pathogenesis of DAMS indicated in the previous research [60]. Also, the main herb modifications for DAMS are the herb pairs like gordon euryale seed and cherokee rose

---

fruit, red poeny root and white peony root, indian bread and largehead atractylodes rhizome, danshen root and fresh rehmannia, etc. These herb pairs are mainly prescribed to treat the accompanying syndromes, such as nephrosis and numbness of extremities, of the DAMS patients.

GPBT syndrome is a rather general syndrome in the patients with the diseases such as chronic gastritis, fatty liver, female infertility, liver cirrhosis and polycystic ovary syndrome. The pathogeneses and manifestations of GPBT syndrome are various. Hence, the herb prescriptions for the patients with GPBT syndrome are also various and individualized. The discovery of the common herb combination knowledge will be helpful to grasp the central rules of the prescriptions for GPBT syndrome. We get 1287 clinical herb prescriptions of GPBT syndrome from the outpatient data of 21 high-experienced TCM physicians in the CDW. There are 367 distinct herbs in the data set, and the result weighted herb combination network has 13,428 edges, which is dense and hard to be directly understood. However, it is interesting that the extracted core herb combinations form a high-experienced classical formula, called Xiaoyao San. Although Xiaoyao San has been accepted as one of the classical formulae (e.g. Xiao Jianzhong decoction, Chaihu Shugan San and Sijunzi decoction) for GPBT syndrome treatment, it has not been suggested as the basic formula for GPBT syndrome treatment yet. Our result proposes a hypothesis that Xiaoyao San could be considered as the fundamental formula to treat the patient with GPBT syndrome. Also, the core herb combinations and the discovered main herb modifications (in the below left part of Fig. 9) provide the individualized therapies for reference in the clinical herb prescriptions for GPBT syndrome. For example, the herb pair of danshen root and red pony root would mainly be prescribed to patients with additional blood stasis syndrome. Based on the analysis result, we have summarized several useful empirical formulae with herb constituents and dosages for the different types of patients with GPBT syndrome [63].

In conclusion, the DM applications based on the TCM CDW get the promising results to help the knowledge distilling and discovery from large-scale clinical data. The results could be a good reference for clinical operations and Chinese patent medicine development. Based on the prepared clinical data from TCM CDW, we will conduct further studies on the various clinical DM topics as indicated in Fig. 8.
6. Conclusion and future work

Clinical research based on the real TCM clinical practices, which keeps to TBSS, is the essential requirement of the development and innovation of TCM. This paper introduces the development of a data warehouse platform for the management, processing and analysis of large-scale SEMR data. We have accomplished the whole framework and developed the core components, such as clinical RIM, ETL tool, OLAP and DM functions. Moreover, based on the collected SEMR data, we have developed and performed several TCM research-oriented subject analyses and DM tasks. The data analysis case studies show that the CDW platform provides a handy approach for TCM clinical knowledge discovery and decision support. Therefore, the CDW will be promising to build an infrastructure for TCM clinical and theoretical research, and promote the development of TCM from individualized empirical summarizing to large-scale evidence-based scientific medical discovery. Meanwhile, the study shows that it is the knowledge management and artificial intelligence approaches, such as clinical reference information modeling, terminology systems, large-scale data integration and preprocessing, and novel DM methods and applications that constitute the fundamental components of the TCM CDW framework. Therefore, we also have developed a practice platform for the innovative study on artificial intelligence approaches and applications in TCM domain. Thus, although the related project of TCM CDW has been initiated since the beginning of 2002, it is still in the preliminary stage and in progress. Currently, the clinical data only contain the TCM research-oriented information, while hospital management information is not included yet. Due to the decision support requirement of hospital management, there needs integrating the data from hospital information system and developing the corresponding subject analyses. However, due to the huge storage requirement and relative independent clinical use of medical image data, we will not consider directly storing the medical image data in TCM CDW. Therefore, we will mainly focus on conducting the following tasks in the future.

The privacy and security issues are the main problems in clinical data sharing and mining. We will address the information content protection on both physicians and patients. This has been partly considered in the current ETL tool (e.g. automatic data anonymization), and the special medical data analysis applications, such as privacy preserving DM methods, should be further studied.

Due to the complicated and unique characteristics of TCM clinical data, we will enhance the clinical information modeling, data preprocessing and DM with the terminological system support. The multidimensional, multi-relational and multi-granular clinical information has produces a real obstacle for DM applications. Most of the variables in TCM clinical data are interdependent from each other, and generated from different information sources or captured by different methods. Hence, the TCM clinical data may have a conflict with the independent and identically distributed assumption. Therefore, developing the novel data modeling and analysis methods will be the further crucial research tasks to explore the TCM clinical data, because there exist more intricate intra- or intra-correlations in the TCM clinical data than in general data.

Compared with the free-text EMR data collection tasks, the collection of high quality SEMR data is still a laborious job. Therefore, the limited data have not made full use of the whole CDW framework. We have hammered at the upgrading of the SEMR system to facilitate the data entry tasks. Furthermore, with more TCM hospitals taking the SEMR system as the regular EMR collecting tool, and more research projects permitted to provide their data, the current data capacity will increase rapidly in the near future.

Acknowledgements

This work is partially supported by Program of Beijing Municipal S&T Commission, China (D08050703020804), China Postdoctoral Science Foundation (2005037106), China Key Technologies R&D Programme (2007BA110806-01), China 973 project (2006CB504601), China NSFC project (90709006) and the S&T Foundation of Beijing Jiaotong University (2007RC072).

References


X. Zhou et al. / Artificial Intelligence in Medicine 48 (2010) 139–152 151

Acknowledgements

This work is partially supported by Program of Beijing Municipal S&T Commission, China (D08050703020804), China Postdoctoral Science Foundation (2005037106), China Key Technologies R&D Programme (2007BA110806-01), China 973 project (2006CB504601), China NSFC project (90709006) and the S&T Foundation of Beijing Jiaotong University (2007RC072).

References


