Data processing and analysis in real-world traditional Chinese medicine clinical data: challenges and approaches

Baoyan Liu,\textsuperscript{a} Xuezhong Zhou,\textsuperscript{b}\textsuperscript{,}\textsuperscript{*}† Yinhui Wang,\textsuperscript{c} Jingqing Hu,\textsuperscript{c} Liyun He,\textsuperscript{d} Runshun Zhang,\textsuperscript{c} Shibo Chen\textsuperscript{c} and Yufeng Guo\textsuperscript{c}

Traditional Chinese medicine (TCM) is a clinical-based discipline in which real-world clinical practice plays a significant role for both the development of clinical therapy and theoretical research. The large-scale clinical data generated during the daily clinical operations of TCM provide a highly valuable knowledge source for clinical decision making. Secondary analysis of these data would be a vital task for TCM clinical studies before the randomised controlled trials are conducted. In this article, we discuss the challenges and issues, such as structured data curation, data preprocessing and quality, large-scale data management and complex data analysis requirements, in the data processing and analysis of real-world TCM clinical data. Further, we also discuss related state-of-the-art research and solutions in China. We have shown that the clinical data warehouse based on the collection of structured electronic medical record data and clinical terminology would be a promising approach for generating clinical hypotheses and helping the discovery of clinical knowledge from large-scale real-world TCM clinical data. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: traditional Chinese medicine; real-world clinical data; data processing and analysis

1. Introduction

Traditional Chinese medicine (TCM) has a unique knowledge system and long history of clinical practice [1]. The general theoretical frameworks based on the ancient Chinese philosophy of regulating the balance of the whole human body system and the clinical prescriptions with multiple components have been the main characteristics of TCM, which are different from conventional biomedicine. Because of the complicated clinical conditions and individualised therapies used in TCM, empirical knowledge accumulated during clinical practice plays a key role in TCM research and theoretical refinement [2]. Although randomised controlled trial (RCT) is considered the best evidence for providing efficacy of healthcare interventions [3], performing TCM RCT is still a real challenge for clinical researchers because of the complicated information involved in diagnosis and treatment. In fact, clinical data and literature accumulated from its daily practice (e.g. inpatient and outpatient encounters) have high values for TCM research. Thus, data processing and analysis, second data analysis in particular, of the observational TCM clinical data, which can unravel hidden patterns or regularities from large-scale clinical records, would be very useful for clinical decision making.

Clinical events (e.g. observation, diagnostic and therapeutic procedures) and the relationships between these events in the clinical data are the significant information components of medical knowledge [4]. Extraction and validation of these types of knowledge would be the main focus of clinical data processing. The main data components with high values in TCM clinical practice also consist of clinical events.

\textsuperscript{a}China Academy of Chinese Medical Sciences, Beijing 100700, China
\textsuperscript{b}School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
\textsuperscript{c}Guang’anmen Hospital, China Academy of Chinese Medical Sciences, Beijing 100053, China
\textsuperscript{d}Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China
\textsuperscript{*}Correspondence to: Xuezhong Zhou, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China.
\textsuperscript{†}E-mail: xzzhou@bjtu.edu.cn

Therefore, studies on how to represent clinical events and how to analyse the relationships between these different events and medical entities (e.g. herb, syndrome and symptom) are essential tasks for TCM clinical data processing and analysis. Medical record is the main way of recording clinical events in TCM, which is represented by the Chinese natural language. Therefore, structured transformation of TCM clinical data is one of the first steps to make them flexible to use. In addition, systemised terminology systems should be incorporated to standardise structured clinical data. Because of high dimensionality and heterogeneous information elements in observational TCM clinical data, large-scale data is needed to get a reliable analysis result. Therefore, it needs large data management techniques, such as a data warehouse, to integrate distributed data sources from different clinical information systems. Furthermore, according to the research topics of TCM clinical studies, clinical data should be filtered and transformed into conventional structures, which are suitable for statistical analysis and traditional data mining methods.

The rest of the article is arranged as follows. We discuss the challenges, such as structured clinical data curation, data preprocessing, large-scale data management and data analysis with complex structures, of TCM clinical data processing and analysis in Section 2. The current status of data processing and analysis is presented in Section 3. Furthermore, the promising approach that we have experienced is introduced in Section 4. Finally, we have a discussion and conclusion in Section 5.

2. Challenges of TCM clinical data processing and analysis

To use large-scale real-world TCM clinical data to support decision making, we should address several technical or methodological obstacles. In this section, related challenging issues will be discussed.

2.1. Structured clinical data curation

Because of the labour-intensive task of clinical practice in daily TCM encounters, particularly outpatient encounters, it is still difficult to collect data from TCM clinics and use it for secondary data analysis. Although current hospitals in China often use the electronic medical record (EMR) system in clinical practice, most of them only support free-text data entry. This means that the main components of the medical record, such as main complaints, medical histories, progress notes and medications, are represented as natural language and in free-text format. To utilise clinical data (e.g. symptoms, herbs and diagnoses) from the medical record, we should first use an approach for recognising medical entities and values from the text. Therefore, structured clinical data curation is needed to help perform the task. Text mining (the method of automatically identifying medically named entities and related values from the textual data) would be one of the useful approaches [6, 7]. However, the structured EMR system that supports the structured input of clinical data would be necessary for higher data quality requirement. A flexibly structured narrative is preferred in electronic health record data capture to balance the rigid structured data entry (SDE) and free-text entry [8]. For SDE, how to get high-quality structured data in the context of a labour-intensive clinical operation environment is a real challenge for TCM practitioners. Because most TCM practitioners would encounter patients with different diseases, capturing symptoms and signs is important for them to diagnose and suggest treatment. Therefore, it is important to integrate symptoms-based clinical terminology systems, such as the International Classification of Primary Care-2 [9, 10], into structured EMR systems to facilitate high-quality data entry in a clinical setting.

In a real-world TCM clinical setting, particularly an outpatient clinical setting, there is often a lack of outcome-related information for further effective data analysis. Without outcome-related information, useful knowledge of diagnosis and treatment cannot be distilled from the empirical clinical data. Therefore, for TCM clinical studies based on real-world clinical data, it is important to include outcome related information, such as health related quality of life [11], disease-specific scales [12, 13] and patient-based measurements [14], in the SDE system.

2.2. Data quality and data preprocessing

Data quality is a fundamental issue in secondary data analysis [15]. Real-world TCM clinical data is recorded for purposes of clinical operation. It contains the necessary information (e.g. symptom, diagnosis, prescriptions and laboratory test results) that is relevant to the diagnosis and treatment of diseases. Other than the survey data used in secondary data analysis [16], which are originally collected for some specific types of clinical research, TCM clinical data include most clinical variables and clinical information valuable for the whole clinical task, and thus could support responses to various clinical
questions, such as diagnosis differentiation, disease epidemiology, herb interactions and optimal clinical therapies [17].

The primary issue in the use of TCM clinical data for analysis is the existence of missing values and heterogeneous data representations. In general, real-world clinical data would usually only record abnormal values of the medical variables of patients. The negative variables are only recorded in the conditions of diagnosis differentiation. Thus, most of the missing values could be considered as negative values. However, because of the personalised views of TCM physicians, some abnormal values would not be recorded in the clinical data if they were considered as having no value for diagnosis and treatment. Therefore, the problem of missing values challenges reliable data analysis. It is important to determine whether the unrecorded positive status of a symptom means that the status was negative or that the status was unknown [18]. Because the missing patterns of clinical data are highly relevant to different TCM physicians, this indicates that physician name or code should be included as one of the essential fields in TCM clinical data, which would be valuable for data analysis.

Furthermore, because of the significant role of ancient TCM textbooks for practical clinical reference [6] and the various terminologies and phrases used in heterogeneous data sources, the same concepts may be represented by different synonymous terms from different ancient TCM textbooks. Moreover, related but not identical phrases are frequently used in the data. Furthermore, data accuracy has become an important concept with the widespread use of computer-based systems for clinical and epidemiological research [19]. Thus, with the structured clinical data prepared, there still needs both a reliable and efficient preprocessing step to integrate and standardise useful data sets from large-scale clinical data sources to provide high-quality data for further analysis.

2.3. Large-scale data management

Traditional Chinese medicine clinical practice incorporates all aspects of TCM clinical information, such as manifestation, diagnosis and treatment. Clinical data includes medical information from both inpatient and outpatient encounters. Furthermore, it is common to use integrated traditional Chinese and conventional medicine in inpatient encounters in TCM hospitals. This means that TCM clinical data would include both TCM variables (e.g. variable of symptom and sign, and herb prescription) and conventional medical variables (e.g. variables of laboratory test results). Therefore, a huge data storage framework is needed to hold clinical data for query and analysis.

Furthermore, because the terms in clinical data from heterogeneous data sources are diverse, it is important to develop a large-scale data management platform with terminology sources to integrate and uniformly organise clinical data and support high performance data query and analysis. To manage large-scale clinical data for efficient data analysis, it would be necessary to design a unified clinical information model to accommodate the different information elements and the complicated relationships between them. Moreover, clinical terminology systems or controlled vocabularies should be related to the clinical data to improve the data quality and usability.

2.4. Multirelational data analysis with rich domain knowledge

Mining over medical, health or clinical data is considered the most difficult domain for data mining [20, 21]. In contrast to the quantitative variables (e.g. laboratory test results) of conventional clinical data, the main contents (e.g. main complaints, histories and diagnoses) of TCM clinical data are originally expressed in the natural language. Moreover, other than conventional medicine, TCM diagnosis and prescription are mainly based on utilising symptom information [22]. Meanwhile, on the basis of TCM theories, TCM has a very elaborate way of differentiating clinical manifestations in patients to make a diagnosis and prescription. Hence, different information components (e.g. manifestations, diagnoses and prescriptions) in structured TCM clinical data are related to each other in a number of ways. Furthermore, synonymous and polysemous terms are difficult to distinguish in structured schema. Hence, TCM clinical data with multidimensional and hierarchical structures form a real challenge for both statistical analysis and data mining research. The relational statistical data analysis and learning methods [23], such as probabilistic relational models [24], relational Markov model and author–topic model [25], could be appropriate methods to explore the related clinical data set. Zhang et al. [26] had proposed a hierarchical author–topic model to find the multilevel symptom–herb associations in 3238 inpatient cases with type 2 diabetes. Using Gibbs sampling algorithm and latent Dirichlet allocation method [27], the model could extract clinically meaningful hierarchical topics including both symptoms and herbs.
from the TCM clinical data set. Given some symptoms of a patient, the model could also predict the corresponding prescribed herbs.

3. Related work and promising approaches

Related data mining research in the TCM field includes syndrome differentiation, syndrome epidemiology and herb combinations, which has been reviewed in several articles [28, 29]. However, previous studies lack standardised terminology and the sample sizes of the studies are usually small. The lack of standardised terminology makes the results of the related studies less comparable and generalised. Furthermore, the results are not reliable when drawn from small-scale data because of the complex information components involved in a real-world TCM clinical setting. There are remarkable differences between a conventional clinical setting and a TCM clinical setting. Based on the macrolevel principles of TCM, the real-world procedure of TCM diagnosis and treatment covers high dimensional factors, such as morbid conditions, natural factors and social factors. These high dimensional factors form huge clinical information space. For example, we have developed a clinical data warehouse [17] with about 40,000 cases, in which there are about 9000 symptom variables and 1000 herb variables. This would create an information space with $10^3$ variables. Thus, it is important to recognise that there is a need for large-scale data in TCM observational studies to generate reliable clinical hypotheses or knowledge. To support large-scale data analysis projects, it is necessary to develop a technique platform to facilitate data curation and processing. It is important to develop a new TCM clinical research framework that focuses on the acquisition, management and analysis of TCM clinical data [2].

3.1. Flexible structured data collection based on clinical terminology

A structured TCM EMR system [30] like OpenSDE (Open Structured Data Entry) [31] was developed in 2002, which helps collect structured EMR data from daily clinical operations. The system supports TCM practitioners with the recording of structured data for use in both practical care and research. Based on flexible recording templates, structured fields with the values of predefined concepts are embedded in free text narratives. Clinicians can print the free text EMR generated by the system for formal medical service and can also use the structured data for research. By 2007, the structured EMR data entry system had collected 20,000 inpatient data with conditions of diabetes, coronary heart disease and stroke from 10 TCM hospitals and TCM wards in the conventional medical hospitals in Beijing. In addition, more than 20,000 outpatient cases from 20 highly experienced TCM physicians in Beijing have been collected to date.

Real-world clinical information is diverse. For example, some patients have the sign of ‘limb pain’, while others have the sign of ‘lower limb pain’. It is clear that the information of ‘limb pain’ covers the information of ‘lower limb pain’. Thus, it is impossible to collect high-quality structured data if comprehensive clinical terminologies are not available. Furthermore, there needs to be a systemised clinical terminology in TCM like SNOMED-CT (Systematized Nomenclature of Medicine–Clinical Terms) [32] with hierarchical concept structures to be used as a reference in the data entry process. To support structured TCM clinical data entry tasks, a systemised TCM clinical terminology [33] has been developed with a terminology classification framework corresponding to the content of TCM EMRs. This terminology system has gathered over 160,000 clinical terms. Currently, the terminology system only includes Chinese terms because the original purpose of the system was to support TCM clinical SDE in China. The extension to other languages like English is necessary for international use and comparable data analysis on a global scope. An SDE system integrating terminologies would be a promising approach for large-scale structured TCM clinical data curation. Furthermore, there is a study indicating that exposure to the computer-based patient record was associated with changes in the information gathering and reasoning strategies of physicians [34]. Therefore, the application and pattern of a structured EMR system used in TCM to improve clinical facility should be further explored.

3.2. Building a clinical data warehouse

A data warehouse [35] is a technical solution for decision support with immense data storage, management and processing. Clinical data warehouses aim to integrate disparate clinical data sources and construct a subject-oriented large-scale data set for clinical decision support, such as hospital infection control [36] and data mining. Zhou et al. [17] introduced the development of a data warehouse platform for the management, processing and analysis of large-scale structured TCM EMR data. They have completed the whole framework and developed the core components, such as clinical reference information...
model, extraction–transformation–loading tool, online analytical processing and data mining functions. Moreover, based on the stored structured EMR data, they have developed and performed several TCM research-oriented subject analyses and data mining tasks. The data analysis case studies [37] show that the clinical data warehouse platform provides a useful approach for TCM clinical knowledge discovery and decision support. Therefore, the clinical data warehouse might have a promising role in the building of an infrastructure for TCM clinical and theoretical research, and provide large-scale high quality real-world TCM clinical data for evidence-based scientific and medical discovery.

3.3. Effectiveness-driven clinical knowledge discovery

Finding clinical regularities of diagnosis and treatment with good outcome is important in promoting the development of sustainable clinical solutions. Currently, information on effectiveness of treatment is often absent in most real-world TCM clinical data. Therefore, most related TCM clinical data mining tasks only generate empirical regularities with no constraints on clinical effectiveness. This means that it is difficult to judge which discovered regularities are effective for disease treatment. For example, association rule mining on herb combinations would generate rules like: if herb A is used, then herb B will be used with a probability of 100%. However, the rule does not provide predictive information on the outcome of patients who have taken both herb A and herb B. Therefore, it is significant to explore clinical knowledge or hypothesis in the context of clinical effectiveness. Furthermore, because TCM theories are rather abstract and only provide the general principle for clinical operations, effective clinical treatment will largely depend on the empirical knowledge of TCM physicians. Therefore, it is promising to discover TCM clinical knowledge by integrating the clinical effectiveness analysis. For example, finding the effective herb–herb interactions from clinical prescriptions for specific disease conditions [38] would be useful to propose regularities for practical clinical treatment. In particular, real-world TCM clinical cases include individualised sequential treatment decision policies. If we can infer optimal or hypo-optimal sequential solutions with effectiveness from the clinical data, it will be a great help for TCM clinical trials and practical clinical treatments. Markovian decision processes (MDPs) and partially observable MDPs [39] are powerful mathematical frameworks for sequential planning under uncertainty and are potential methods to help plan the sequential treatment for specific diseases [40]. Feng et al. [41] proposed an initial study on using MDP to model traditional Chinese medicine therapy planning to find the optimal sequential clinical treatment options for inpatients with type 2 diabetes. It inferred the model parameters (e.g. states, actions, rewards and transition functions) from clinical data set and used value iteration algorithm [42] to solve the MDP problem. The study shows that MDP model can be used to help identify useful clinical prescriptions from data and help to generate sequential TCM clinical guidelines for type 2 diabetes treatments.

3.4. Text mining from free-text EMR data

Because SDE is an additional labour-intensive task for TCM practitioners, structured data curation performed directly by TCM practitioners needs continuous training and practice. Previously used free-text EMR systems in China have generated large-scale EMR documents. Thus, it would be helpful to use an automatic tool for extracting structured data from free-text EMR documents. Text mining methods, such as information extraction and named entity recognition, could be used for this purpose. Although dissecting sentences and extracting the meaning completely from free-text clinical documents is still difficult [43], text mining tools are becoming increasingly used and valued for simplified tasks [44–46]. Because of the heterogeneity of free-text EMR documents generated by different TCM practitioners and the need for word segmentation tasks in Chinese language texts [6], it is even more difficult to perform text mining in free-text TCM EMR data than in free-text conventional medical EMR data. Thus, to improve the quality of text mining results, systemised clinical terminology systems including synonyms and semantic relationships, such as UTCMLS [47] and TCM clinical terminology [33] would be important to support ontology-based information extraction from TCM clinical texts.

4. Discussion and conclusion

Real-world clinical practices play a significant role in TCM research. Because of its individualised nature and ad hoc prescriptions for different patients in TCM clinics [48], large-scale real-world TCM clinical data consisting of novel regularities generated by TCM practitioners are a highly valuable data source.
for research. Therefore, data analysis and processing of real-world clinical data would be an important step in transforming individualised empirical regularities into high-level medical evidences.

Evidence-based medicine has been recognised as an important method in TCM clinical studies [49]. There is a large amount of case reports and observation studies in the TCM field. Using statistical analysis methods, the results from previous studies form a good reference for TCM practitioners. However, because of the large information space held in TCM data, most of the related observational studies have sample sizes that are too small to get reliable results. Furthermore, a study [50] shows that RCTs in the TCM field are still relatively low in quality. This may be attributed to the improper reduction of the complicated and large amount of medical factors involved in real-world TCM diagnosis and treatment. Therefore, it is both a significant and promising approach to develop a large-scale clinical data warehouse based on the structured high-quality data extracted or curated from daily clinical practice. When large-scale real-world clinical data in the TCM field is prepared by using data processing and analysis methods, it would help produce large numbers of clinical hypotheses, which would be good preparation for high quality TCM RCTs.

Recently, increasing interest has been shown in using routinely collected clinical data (mainly real-world EMR data) for medical studies [51–53]. However, it is still difficult to process real-world clinical data because of the complicated and diverse information elements involved and data quality control issues. Multirelational clinical data pose a real challenge for statistical analysis and data mining methods [53]. To mine the data from multiple relations and extract regularities directly from databases without having to flatten data led to the emergence of an active new field of research: multirelational data mining (MRDM) [54]. MRDM in the biomedical field has been one of the most active fields because of the rich relational structures incorporated in large-scale data [55]. Real-world TCM clinical data analysis would be another potential field for MRDM research. As a real-world clinical data capture procedure performed in daily clinical practice, it has both advantages and disadvantages for data quality control. Real-world clinical data capture does not need worry about the authenticity of data because the corresponding EMR data should keep to legal duty. Therefore, the reliability of real-world clinical data analysis would mainly attribute to data selection from large-scale clinical repository and normalization of terminological data in clinical data set. This means that most data quality control efforts and methods should focus on the preprocessing of collected real-world clinical data to get a high quality data set for analysis. Meanwhile, to generate the substantial clinical solutions for real-world patients, it is significant to collect clinical effectiveness information in each clinical case and explore those methods incorporating clinical effectiveness related variables (e.g. partially observable MDP) to find the practically effective treatment options.

Acknowledgements

This work was partially supported by the China 973 project (2006CB504601), Program of Beijing Municipal S&T Commission, China (D08050703020803 and D08050703020804), China Key Technologies R&D Program (2007BA110B06-01), NSFC Project (61105055) and National S&T Major Project of China (2009ZX10005-019 and 2009ZX09301-005-010).

References


Artificial Intelligence in Medicine


ACM SIGKDD


