

## RESEARCH ARTICLE

# Understanding traditional Chinese medicine via statistical learning of expert-specific Electronic Medical Records

Yang Yang<sup>1,†</sup>, Qi Li<sup>1,†</sup>, Zhaoyang Liu<sup>1</sup>, Fang Ye<sup>2</sup>, Ke Deng<sup>1,\*</sup>

<sup>1</sup> Center for Statistical Science & Department of Industry Engineering, Tsinghua University, Beijing 100084, China

<sup>2</sup> Zhou Zhongying's Studio, Nanjing University of Chinese Medicine, Nanjing 210046, China

\* Correspondence: kdeng@tsinghua.edu.cn

Received August 28, 2018; Revised January 16, 2019; Accepted March 26, 2019

**Background:** Traditional Chinese medicine (TCM) has been attracting lots of attentions from various disciplines recently. However, TCM is still mysterious because of its unique philosophy and theoretical thinking. Due to the lack of high quality data, understanding TCM thoroughly faces critical challenges. In this study, we introduce the Zhou Archive, a large-scale database of expert-specific Electronic Medical Records containing information about 73,000 + visits to one TCM doctor for over 35 years. Covering the full spectrum of diagnosis-treatment model behind TCM practice, the archive provides an opportunity to understand TCM from the data-driven perspective.

**Methods:** Processing the text data in the archive via a series of data processing steps, we transformed the semi-structured EMRs in the archive to a well-structured feature table. Based on the structured feature table obtained, a series of statistical analyses are implemented to learn principles of TCM clinical practice from the archive, including correlation analysis, enrichment analysis, embedding analysis and association pattern discovery.

**Results:** A structured feature table of 14,000 + features is generated at the end of the proposed data processing procedure, with a feature codebook, a term dictionary and a term-feature map as byproducts. Statistical analysis of the feature table reveals underlying principles about the diagnosis-treatment model of TCM, helping us better understand the TCM practice from a data-driven perspective.

**Conclusion:** Expert-specific EMRs provide opportunities to understand TCM from the data-driven perspective. Taking advantage of recent progresses on NLP for Chinese, we can process a large number of TCM EMRs efficiently to gain insights via statistical analysis.

**Keywords:** TCM; EMRs; data-driven perspective; Chinese text mining; statistical analysis

**Author summary:** Traditional Chinese medicine (TCM) is attracting more and more attentions from various disciplines. But TCM is still mysterious due to its unique philosophy, model and theoretical thinking. In this paper, we introduce the Zhou Archive, a large-scale database of expert-specific Electronic Medical Records (EMRs) containing visits to one TCM doctor. We transform the original EMRs into a well-structured feature table by multiple data processing tools. Based on this structured feature table, a series of statistical analyses are implemented to learn principles of TCM clinical practice, which reveal insights to understand TCM from a data-driven perspective.

## INTRODUCTION

Traditional Chinese medicine (TCM) has a long history of over 2,000 years, and once played an important role in

healthcare in pre-modern East Asia. As an important branch of alternative medicine, it has been becoming more and more popular worldwide in recent years, and attracting a lot of attentions from scientists of various

<sup>†</sup> These authors contributed equally to this work.

disciplines. For example, Refs. [1–3] confirmed the unique treatment effects of acupuncture; Refs. [4–6] provided insights on how TCM prescriptions work via systematic interactions with biological regulation network; and, the 2015 Nobel Prize awarded to Prof. Yoyo Tu for her contribution to the discovery of artemisinin in 1977 casted lights on the great impact of TCM on human beings.

On the other hand, however, TCM is still mysterious to many people because of the unique philosophy, model and theoretical thinking behind it. Similar to any other healthcare systems, TCM also contains three basic components: (i) a toolbox of therapeutic technologies to treat patients, (ii) biomedical measurement instruments to observe and measure physical status of patients, and (iii) a diagnosis-treatment model (DTM) to map the biomedical observations and measurements of a patient to a “proper” therapy in the toolbox. But, due to the philosophical environment of ancient China and technical constraints in history, TCM developed these components in a unique way.

First, TCM therapies typically have complex internal structures. TCM prescriptions and acupuncture are the two primary therapies of TCM (although there are records of surgeries in the long history of TCM). A TCM prescription typically contains multiple ingredients, which may generate a mixture of hundreds of chemical compounds. An acupuncture therapy is usually composed of a series of acupunctures in different locations (called acupoints) of the patient’s body. The combinatorial or sequential nature of TCM therapies provides flexibility to tune treatment adaptively based on status of patients, but also posts great challenges in quality control and efficacy evaluation of TCM therapies. Second, due to the technology constraints in history, biomedical measurements of TCM heavily depend on subjective observation of doctors, and rely on natural language to deliver the experience. The combination of subjectivity of observations and flexibility of natural language may introduce multiple levels of bias and noise to the measurements, leading to critical technical barriers in data analysis. Third, built on top of the Chinese philosophy, the diagnosis-treatment model of TCM is described in a unique language involved many philosophical concepts in ancient China whose concrete meanings may change over time and be interpreted in different ways. This phenomenon makes it a challenging job to decode and understand the diagnosis-treatment model of TCM from a positive perspective.

All these features shaped TCM into a healthcare system with a unique knowledge representation style and deduction logic, which is very different from the modern healthcare system developed in the western world on top of anatomy and cell/molecular biology. In the past

decades, many efforts have been given to build connections between TCM and modern sciences, trying to evaluate, understand and reinterpret TCM in a modern way. These efforts can be roughly classified into two categories: (i) the drug-discovery oriented research, which aims to identify potential drug candidates and validate them via randomized experiments [7–11]; and (ii) the theory-understanding oriented research, which focuses on revealing causal mechanism or association patterns of the diagnosis-treatment model behind TCM via data-driven approaches [4,12–18]. Although there are many difficult issues in practical implementation, the drug-discovery oriented research enjoys a relatively straightforward logic. The theory-understanding oriented research, however, often faces critical challenges at both methodology level and data level.

At the methodology level, it is very challenging to design data models that can precisely reflect TCM thinking and/or appropriately approximate generating procedure of TCM data. At the data level, a major problem is the lack of high quality data carrying stable signals about the full spectrum of TCM clinic practice. It’s not difficult to find a small-scale dataset with hundreds or thousands of patients from one TCM doctor. But, such a dataset is often biased to a small patient population of a certain disease. It’s also possible to assemble many small-scale datasets into a large-scale dataset. But, a dataset generated in this way is often a mixture of many inconsistent components, leaving many uncontrollable risks in downstream data analysis.

In this study, we introduce the Zhou Archive, a large-scale database of expert-specific Electronic Medical Records (EMRs), which contains comprehensive information about 73,000+ visits to one TCM doctor by 26,000+ distinct patients over 35 years from 1980 to 2015. From many perspectives, the archive provides an ideal opportunity to understand TCM in a data-driven way. First, the scale of archive is large enough to support many data-driven approaches. Second, the 73,000+ visits by 26,000+ patients cover 1,300+ diseases of 16 major disease categories, including cancers, digestive diseases, infectious diseases, neurological diseases, respiratory diseases, cardiovascular diseases, urinary diseases, rheumatism and so on, and are rich enough to reflect all aspects of TCM practice. Third, with data fields for symptoms of patients, TCM diagnosis and TCM treatment, the archive records all key components of the diagnosis-treatment model behind TCM, making it possible to decode the model in a data-driven way. Moreover, as all EMRs in the records come from one TCM doctor alone, the underlying logic of diagnosis-treatment model is more likely to be self-consistent, which is extremely important to the success of data-driven approaches. At last, except for classic TCM features, the

archive also contains information about lab tests and diagnosis from the western medicine perspective, allowing us to connect TCM concepts with Western Medicine.

With the rise of medical big data and the popularity of precise medicine in recent years, real world study based on large scale EMRs has become an important paradigm in healthcare research [19–25]. We hope this study can open a door to this paradigm for TCM-related studies. Like most EMR data in practice, the data in the archive is a mixture of structured data fields which encode information with a well-design feature table, and semi-structured/unstructured data fields which deliver information via semi-structured or free texts. To transform the original EMR data into a well-structured feature table for which statistical analysis can be implemented, we need to discover a lot of TCM-specific and archive-specific technical terms from the archive, map them to their standard feature codes, and properly process the semi-structured and free Chinese texts in the archive to decode information effectively. In this paper, we proposed a systematic data processing framework to achieve this goal.

Based on the structured feature table obtained, a series of statistical analyses are implemented to learn principles of TCM clinical practice from the archive. Cross-category association patterns are discovered using various technical tools and embedding analysis is used on prescriptions and symptoms. Results from these analyses reveal insights to understand TCM from a data-driven perspective.

The remainder of this paper is organized as follows. “Description Of The Data” briefly introduces the data structure of the Zhou Archive. “Transferring Semi-Structured EMRs Into A Structured Feature Table” proposes a data processing framework to transform the original semi-structured and unstructured data from the archive to a well-structured feature table. In “Statistical Learning Of The Structured Feature Table”, we analyze the structured feature table obtained with a series of statistical methods and extract some hidden patterns of this database. Finally, we summarize and discuss this study in the last section.

## DESCRIPTION OF THE DATA

The archived EMRs contain 14 distinct data fields of 6 categories, including: (i) Patient ID and Demographics (ID, Gender, Age), (ii) Visit Date, (iii) Clinical Features (Symptoms, Tongue Picture, Pulse Type, Lab Tests), (iv) Western Medicine Diagnosis (Disease, Disease Category), (v) TCM Diagnosis (TCM Disease, TCM Pathogenesis) and (vi) TCM Treatments (TCM Therapy, TCM Prescription).

The 14 data fields can be classified into three types: 7

structured fields encoding information with well-designed codes (including Patient ID, Gender, Age, Visit Date, Disease, Disease Category, TCM Disease), 6 semi-structured fields encoding information with semi-structured texts (including Tongue Picture, Pulse Type, Lab Tests, TCM Pathogenesis, TCM Therapy, TCM Prescription), and 1 unstructured field that delivers information with free texts (*i.e.*, Symptoms). All these data fields contain missing values.

In the database, the column “Western Medicine Diagnosis” comes from the records of visiting western medicine doctors before coming to Prof. Zhou. These western medicine diagnoses were recorded by Prof. Zhou in the archive, each for one visit. Totally, 1,339 distinct diseases appear in the archive, which can be further classified into 16 disease categories, including: Cancers, Digestive Diseases (DD), Infectious Diseases (InD), Neurological Diseases (ND), Respiratory Diseases (RD), Cardiovascular Diseases (CD), Urinary Diseases (UD), Rheumatism, Gynopathy, Skin Diseases (SD), Hematopathy, Endocrine Diseases (ED), Orthopedic Diseases (OD), Ophthalmological and Otorhinolaryngological Diseases (OOD), Men Diseases (MD), and Miscellaneous Diseases (MiD). In terms of TCM diseases, however, only 394 distinct TCM diseases appear, partially due to the higher missing rate of the TCM Disease field (88.4%) than the Disease field (17.5%). More detailed information about the patients covered by the archive is provided in Supplementary Figure S1A–S1E.

One third of the patients in the archive visited Prof. Zhou for multiple times. These patients with longitudinal records paid 4.7 visits on average within an average time span of 242 days, and the average time gap between two adjacent visits is 65 days. Supplementary Figure S1F–S1H give the detailed distributions of visit frequency, overall time span and time gap between two adjacent visits of these patients. Researchers who are interested in this archive can check the website of Zhou Archive for TCM Study for detailed information on the data structure and data access.

## TRANSFERRING SEMI-STRUCTURED EMRs INTO A STRUCTURED FEATURE TABLE

With both structured data fields and semi-structured/unstructured data fields, the original archive is difficult to analyze. In this section, we transform the semi-structured and unstructured EMRs of the archive into a well-structured feature table, for which statistical analysis can be conveniently implemented. To achieve this goal, we process the structured fields, semi-structured fields and unstructured Symptoms field separately by different data

processing strategies.

Figure 1 shows the route map of the data processing procedure, which digests the original archive as the input, and returns the following outputs: (i) a feature codebook  $F$  which encodes all features generated from the archive, (ii) a term dictionary  $D$  which fully covers the vocabulary specific to the archive (including all background words, common TCM terms and special terms used by Prof. Zhou), (iii) a term-feature map  $M$  which links terms in  $D$  and the standard feature codes they correspond to, and the most importantly, (iv) a well-organized structured feature table  $T$  with columns for different features and rows for different records. Different from the raw data in the archive, which delivers information via semi-structured and unstructured texts, the transformed two-dimensional feature table  $T$  encodes information with a well-designed data format and coding system.

There are a few critical challenges in this data processing procedure due to the semi-structured and unstructured texts in the archive. First, text segmentation and term discovery. As there are no visible word

boundaries such as spaces in Chinese texts, the unstructured Chinese texts in the Symptoms field must be segmented into sequences of meaningful terms to decode information. However, because these texts contain many domain specific words, phrases and technical terms that are previously unknown, text segmentation is entangled with term discovery in this study. The combination of these two critical problems posts great challenges in processing the free texts in the archive. Second, standardization of technical terms. Due to the flexibility of free texts, many technical terms in the archive have multiple variates. To sufficiently extract information from the data, we need to map different variates of a technical term to its standard code. Third, we also need to understand the semantic meaning of semi-structured and free texts in the archive to precisely decode information.

Although many tools have been invented to process Chinese texts in the past decades, it is still not trivial to overcome above challenges in this study. Here, we propose an integrative data processing framework as a preliminary solution to this important but challenging

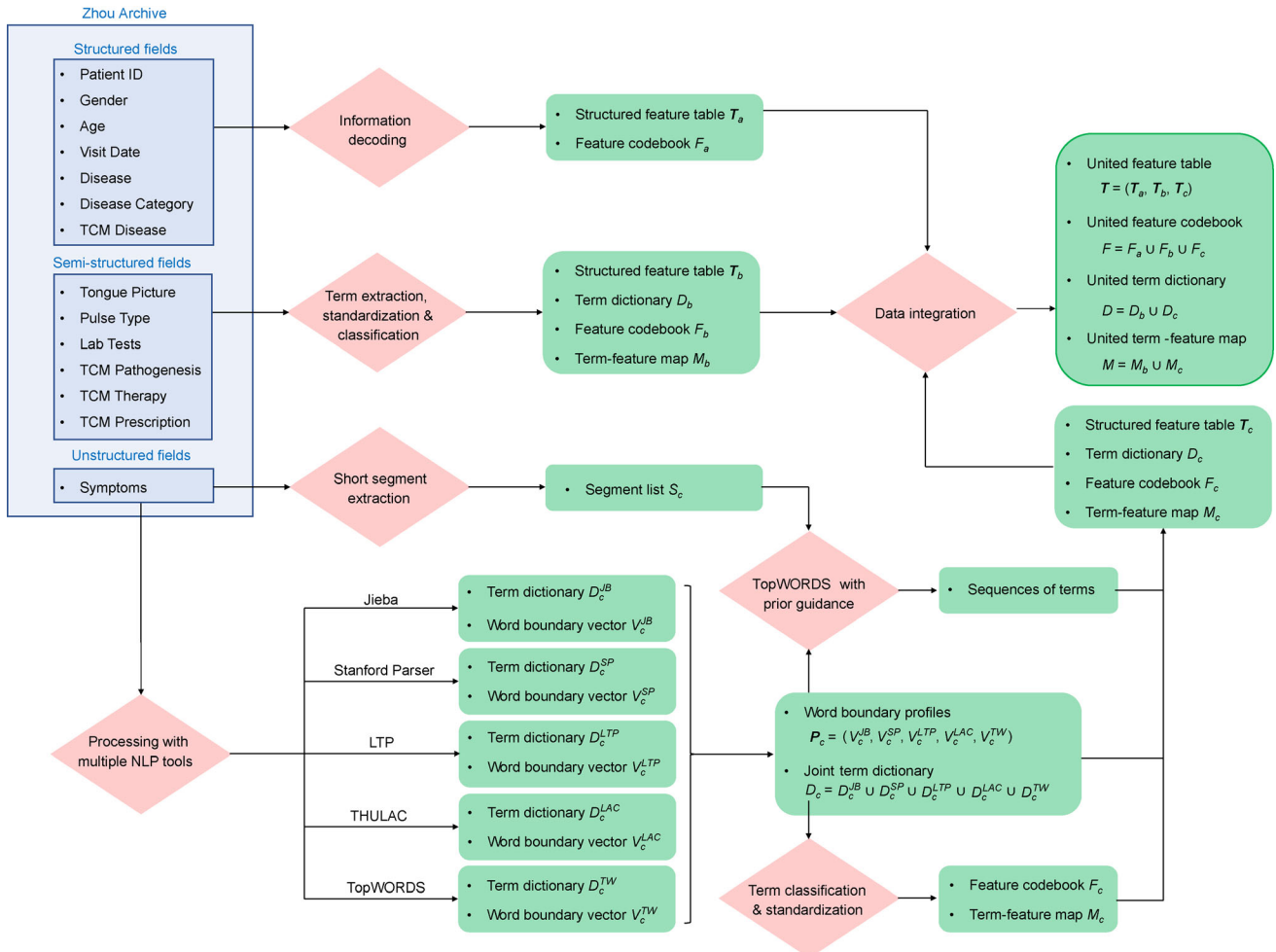


Figure 1. Flowchart to transfer the original Zhou Archive to a structured feature table.

problem. As the same problem will be encountered in many similar studies in the future, we hope that the framework we suggest can serve as a baseline solution for researchers in this field.

### Processing the structured and semi-structured data fields

First, we process the structured and semi-structured data fields, transforming them into a feature table. Because the structured data fields already encode information with a well-designed feature codebook, it is straightforward to decode these fields to get the feature codebook  $F_a$  and a feature table  $T_a$ .

For the semi-structured data fields, however, we need to make extra efforts to collect technical terms in these fields and transform them into their standard feature codes. Taking advantage of the existing data structure in these semi-structured fields, a lot of technical terms can be conveniently extracted. For example, tongue/pulse-related terms and lab tests in the Clinic Feature fields, terms in the TCM Pathogenesis and TCM Therapy fields, as well as herb names in the Prescription field, can be obtained in a straightforward way by enumerating Chinese strings segmented by commas or numbers in the according data fields. Totally, 5,000 + distinct terms are extracted in this way, forming a dictionary of terms denoted as  $D_b$ . Table 1A shows the most frequent terms extracted from each of these semi-structured fields.

These extracted terms need to be transformed to their standard codes before downstream analysis can be proceeded. This can be achieved via two typical operations: *splitting* and *mapping*. Many terms extracted from these semi-structured fields tend to abbreviate multiple concepts to a single term. For example, term “taihouhuang” from the field of Tongue Picture is the abbreviation of two terms “thick tongue fur” and “yellow tongue fur”, term “maixianhua” from the field of Pulse Type is the abbreviation of two terms “stringy pulse” and “slippery pulse”, term “ganshenkuixu” from the field of TCM Pathogenesis is the abbreviation of two terms “deficiency of liver” and “deficiency of kidney”. Standardization of these terms can be achieved by identifying the multiple concepts compressed in one term, and listing the standard feature codes of these concepts in parallel (e.g., “taihouhuang” → “thick tongue fur, yellow tongue fur”). We call this operation as “splitting”, as it divides one technical term into multiple features. On the other hand, many extracted terms refer to the same concept. For example, term “manyigan” is an abbreviation of “chronic hepatitis B”; term “dashengdi” and “xishengdi” refer to the same herb “dried rehmannia root”. Standardization of these terms can be achieved by “mapping”, i.e., building a mapping table from these

terms to their standard feature codes (e.g., “dashengdi” → “dried rehmannia root”, and “xishengdi” → “dried rehmannia root”). Please note that we may need the combination of splitting and mapping sometimes to standardize a term with complex structure.

Totally, 4,000 + features are generated for the 5,000 + extracted terms, resulting in a feature codebook  $F_b$ . The transformation rules from terms in  $D_b$  to features in  $F_b$  are summarized in a term-feature map  $M_b$ , based on which a structured feature table  $T_b$  can be established from the semi-structured fields.

### Unique properties of the free texts in the Archive

Next, we process the free texts in the unstructured Symptoms field. These free texts contain 1,177,007 Chinese character tokens, of which 2,678 are unique. From the text analysis perspective, these texts are unique in multiple dimensions. First, these texts contain a lot of TCM-specific technical terms rarely used elsewhere and many special terms invented by Prof. Zhou that are specific to this archive only. Second, some segments of these texts are highly repetitive. Cutting down these free texts into small pieces separated by natural boundaries (such as punctuation marks, ends of lines and so on), we obtained ~241,000 segments, of which ~126,000 are unique. Most of these unique segments are short strings with  $\leq 10$  Chinese characters, and many of them repeat heavily in the free texts: 3 segments appear  $\geq 1,000$  times, and the 2,100 + segments that appear more than 10 times contribute ~90,000 repeats together, which equivalents to 1/3 of the total number of segments generated from the free texts. We summarize these unique segments into a segment list  $S_c$ . Table 1B shows the top 100 segments with the highest repeat frequency in  $S_c$ . Third, these texts are written in a unique style that is very different from classic training corpus for Chinese text mining, which is typically based on news articles.

These facts mean that we need to capture the special technical terms in the archive to establish an archive-specific vocabulary, and a style-robust tool to process the free Chinese texts in the archive. Moreover, as most segments (especially these highly repeated ones) in the segment list  $S_c$  can deliver one piece of intact information about patients, it is more efficient to achieve semantic understanding with these segments, instead of words or terms, as the basic language units.

### Processing the unstructured symptoms field

In the past decades, many tools for processing Chinese texts have been proposed. In this study, we tried four popular “supervised” methods, namely Jieba, Stanford Parser [26–27], Language Technology Platform (LTP)

**Table 1 Top terms discovered from semi-structured/unstructured data fields**(A) Top 20 terms of the 6 different semi-structured data fields in term dictionary  $D_h$ 

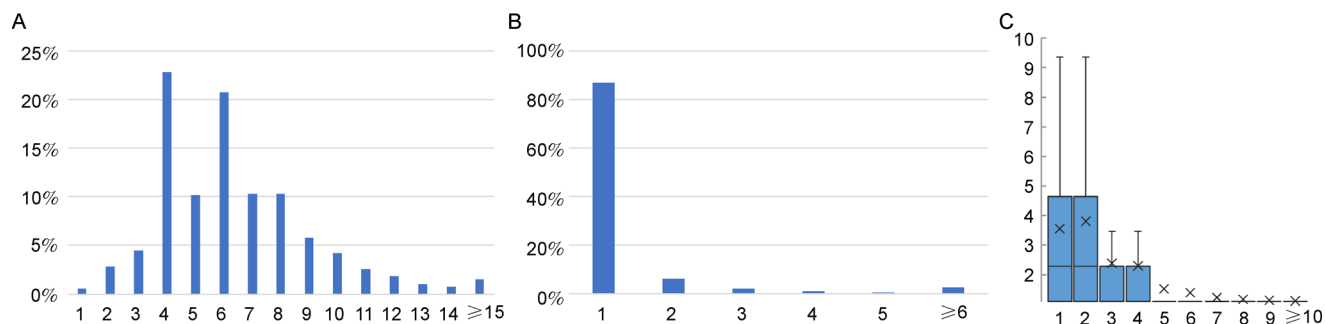
Tongue Picture (503)	Pulse Type (142)	Lab Tests (393)	TCM Pathogenesis (1,033)	TCM Therapy (1,699)	Herbs (1,515)
Dark	Thin pulse	B ultrasound	Impairment of both Qi and Yin	Invigorating Qi and nourishing Yin	Rhizoma pinellinae praeparata
Dark red	Slippery pulse	Biochemical test	Weakness of liver and kidney	Tonifying liver and kidney	Pseudostellariae radix
Yellow, thin and greasy tongue fur	Stringy pulse	CT	Syndrome of liver and stomach disharmony	Regulating liver and spleen	Radix glycyrrhizae
Red	Small pulse	Blood examination	Dampness-heat obstruction syndrome	Clearing dampness-heat	Salvia miltiorrhiza
Thin, yellow and greasy tongue fur	Rapid pulse	Echogastroscope	Kidney deficiency and sthenic liver-energy	Clearing dampness-heat and stasis toxin	Red paeonia
Yellow tongue fur	Soft pulse	Liver function test	Dampness-heat in the interior	Catharsis and thanhhoa	Boiled bombyx batryticatus
Thin and yellow tongue fur	Deep pulse	BP	Deficiency of liver and kidney	Dissipating phlegm and removing blood stasis	Poria cocos
Yellow and greasy tongue fur	Relaxed pulse	ALT	Impairment of both liver and spleen	Strengthening body and anti-cancer	Processed rhizoma
Light yellow, thin and greasy tongue fur	Weak pulse	Fatty liver	Phlegm stagnation in collateral	Invigorating spleen and stomach	Unprocessed rehmannia root
Thin tongue fur	Feeble pulse	Liver function retest	Weakness of both Qi and Yin	Nourishing upper warmer	Coptidis
Dark violet	Left slippery pulse	Urinalysis	Yin deficiency of liver and kidney	Regulating the thoroughfare	Fried atractylodes macrocephala koidz
Reddish	Right slippery pulse	HP	Phlegm and blood stasis	Diffusing and clearing upper warmer	Glehniae radix
Light yellow and greasy tongue fur	Right thin pulse	AST	Spleen deficiency and stomach weakness	Treating both cause and symptoms	Ophiopogonis radix
Cracked	Left stringy pulse	CA199	Wind-phlegm stagnation	Relieving liver and gallbladder	Dried orange peel
Violet	Right stringy pulse	WBC	Combination of dampness-heat and stasis toxin	Activating blood and dredging	Ligusticum wallichii
Toothed	Left thin pulse	Routine urine test	Lower weakness in liver and kidney	Hyperactivity nourishing heart to calm mind	Astragali radix
Thin and greasy tongue fur	Uneven pulse	CEA	Endogenous wind rise	Treating symptoms first	Rhizoma anemarrhenae
Light yellow tongue fur	Irregular pulse	CA125	Impairment of both body fluids and Qi	Nourishing liver and kidney	Caulis spatholobi
Dark and light purple	Left small pulse	HBsAg	Occurrence of cancer toxin	Nourishing blood and thinning liver	Andeophorae radix
Tongue tip red	Right small pulse	MRI	Kidney deficiency and liver depression	Strengthening spleen in transportation	Barbary wolfberry fruit

(B) Top 100 segments with the highest repeat frequency in segment list  $S_c$  (continued)

Term	Count	Term	Count	Term	Count	Term	Count	Term	Count
Dry mouth	3,353	Abdominal distension	482	Insomnia	259	Edema of lower limbs	168	Hidden pain in liver	139
Deep-colored urine	1,448	Belching	461	Feel agitated	256	Headache	168	Numb hand	137
Fatigue	1,158	Pathology	413	Vomiting	230	Dreaminess in night	168	Feverishness in palms and soles	137
Sensation of chill	983	Cough	378	Gasteremphraxis	223	Stiff neck	168	Short of breath	136
Poor sleep	934	Good appetite	371	Hypertension	220	Unnormal stool	166	Cholecystitis	135
Poor appetite	893	Frequent urinary	357	Dreaminess	213	Not painful	165	Color in yellow	135
Dizziness	769	A little dry mouth	355	Out of breath	213	Not a lot stool	162	Down more than quantity	134
Good defecation	763	Backache	331	Yellow face and poor looking	210	Unobvious dry mouth	158	Stomache	133
Oppression in chest	758	Slightly decayed stool	310	Sore throat	209	Dry throat	156	Eat a little	133
Soreness of waist	757	Slightly regular bowel	308	Feeble leg	207	Frequent passing of flatus	156	Emaciation	131
Normal bowel	746	Borborygmus	300	Feel tired	199	Heating palm	152	Fatty liver	131
Regular bowel	658	Sweating	299	Feel fatigue	193	Noisy heart	151	Much poor sleep	131
Bitter taste in mouth	612	Acid regurgitation	290	Inhibited defecation	189	Poor breathing	150	Limb leg	131
Dry mouth and want to drink	568	Tinnitus	288	Normal hepatic region	185	Constious fatigue	149	CT	130
Dry and hard stool	564	Dry stool	286	Sick to vomiting	176	A lot of menstruation	148	Dry mouth at night	128
Sweat easily	562	Undry mouth	286	Not too many self-conscious symptoms	176	Deep sore throat	145	Drink a little	127
Slightly dry stool	533	Taste food well	274	Blurred eye	173	Not much expectoration	145	Normal stool with shape	126
Dry and bitter taste in mouth	528	Nauseating	271	Loose stool	172	Dry and bitter mouth	141	Little in menstruation	125
Palpitation	504	Dizziness	269	Poor looking	172	Odor in mouth	140	Sweaty a lot	125
Normal appetite	492	Not dry stool	264	Quite good appetite	172	Get cold easily	140	Poor appetite	122

(C) Top 30 nontrivial terms of 6 different categories in term dictionary  $D_c$  (continued)

Symptoms (14,346)			Body parts (3,665)		Disease names (2,050)		Lab tests (956)		Medical treatments (437)		Background terms (60,623)	
Dry mouth	Dry stool	Hand	Hand	Gastral cavity	High blood pressure	Gastroitis	Blood pressure	Hemameba	Chemotherapy	Take TCM	Stool	Past days
Cough	Loose stool	Abdomen	Abdomen	Lower limb	Hepatitis B	Myoma of uterus	CT	Protein	Radiotherapy	Gallbladder removal	Pain	Examination
Fatigue	Bitter taste in mouth	Stomach	Brain		Adenocarcinoma	HLP	liver function test	MRI	After chemotherapy	Take prednisone	Surgery	Treatment
Debility	Nausea	Heart	Liver		Gastritis	Cervical spondylosis	health examination	Liver function test	Chemotherapy after surgery	Gallbladder surgery	Normal	Hospital
Vertigo	Poor appetite	Head	Nose		Diabetes	Gastric cancer	Gastroscope	Liver function	Chemotherapy and radiotherapy	4 cycles of chemotherapy	Now	Test
Sensation of chill	Dyspepsodynia	Eye	Lymph		Capsulitis	Enteritis	B ultrasound	Urine test	6 cycles of chemotherapy	After radiotherapy	Obviously	Unwell
Oppression in chest	Headache	Chest	Joint		Cholecystitis	Squamous cancer	Glucose	Stool volume	Take insulin	Gamma knife treatment	Sometimes	Discover
Deep-colored urine	Belching	Waist	Bone		Liver cirrhosis	Rhinitis	Ascites	Enteroscopy	Colon cancer surgery	1 time of chemotherapy	Less	Usually
Fatigue and debility	Dizziness	Foot	Ear		Hypertensive disease	Hepatitis B disease	Fatness in liver	CT scan	Do chemotherapy	2 times of chemotherapy	Self conscious	Transfer
Abdominal distension	Insomnia	Lung	Upper body		Gallstones	Coronary heart disease	Menstrual blood volume	Heart rate	Induced abortion	Western medicine control	Last year	Currently
Palpitation	Gasteremphraxis	Back	Hepatic region		Hepatopathy	Nephritis	Blood volume	Electrocardiogram	Rectal cancer surgery	2 cycles of chemotherapy	This year	Worse
Soreness of waist	Swelling pain	Gallbladder	Hand and foot		Lung cancer	Lipemia	Body weight	Occult blood test	Take antihypertensive drug	Hormone treatment	Disease history	Long time
Dull pain	Numbness	Intestine	Lymph gland		Intestinal cancer	Hyperlipemia	Three positive	ALT	Take western medicine	Take chemotherapy medicine	Recently	Urine and stool
Poor sleep	Catching cold	Gland	Shoulder		Hepatitis	Liver cancer	Test blood pressure	Renal function	TCM treatment	6 times of chemotherapy	Appetite	Left side
Stomachache	Well gas	Neck	Face		Cerebral infarction	Breast cancer	urine volume	Blood type	4 cycles of chemotherapy	Successive chemotherapy	Discomfort	After treatment



**Figure 2. Statistical properties of segments cut from the free texts in the Symptom field.** (A) Segment length. (B) Repeat frequency. (C) Repeat frequency by length

[28] and THU Lexical Analyzer for Chinese (THULAC) [29–30], and a recently proposed “unsupervised” method called TopWORDS [31], to process the free texts in the unstructured Symptoms field.

The four supervised methods emphasize precise text segmentation under the guidance of a preloaded vocabulary and high-quality training corpus. They typically match the target texts with words from a preloaded vocabulary, and do statistical inference when meeting ambiguous words based on a statistical model trained by manually segmented and labelled training corpus. When the actual vocabulary is covered by the preloaded vocabulary and writing style of the target texts are close to the training corpus, these supervised methods usually perform pretty well. But, previous study [31] also showed that when the actual vocabulary of the target texts contains a lot of words beyond the preloaded vocabulary, they often fail to recognize many of these unregistered words, especially when the writing style of the target texts is very different from the training corpus.

The unsupervised method TopWORDS, however, pays more attention on efficient new word discovery, although it can also be used as a tool for text segmentation. It can effectively discover previously unknown words and phrases when no preloaded vocabulary and proper training corpus are available, or the preloaded vocabulary and the training corpus do not fit the target texts well. Detailed information about the five NLP tools for processing Chinese texts can be found in the Appendix.

As shown in Figure 1, each of these methods returns a *term dictionary*  $D$  and a *term boundary vector*  $V$  as the outputs for term discovery and text segmentation, respectively. The term dictionary  $D$  is the set of terms identified by the method, and the term boundary vector  $V$  is a vector with the same length of the target texts, whose element  $V_i$  can take three values:  $V_i = 2$  if there is a

natural boundary (e.g., punctuation mark, end of line and so on) behind the  $i$ th position of the target texts,  $V_i = 1$  if the method puts a term boundary there, and  $V_i = 0$  otherwise. The detailed term boundary vectors of different segmentation tools can be found in the website of “Zhou Archive for TCM Study”. Table 2 summarizes and compares their performance in multiple angles, from which we can see that both the reported term dictionary and the predicted term boundary vector vary significantly across different methods, indicating the critical challenges in processing and understanding domain-specific Chinese texts.

Table 2A summarizes the number of nontrivial terms (i.e., terms with more than one Chinese character) discovered by different methods. The number varies from the smallest 23,989 terms reported by Jieba to the largest 47,248 terms reported by TopWORDS. Ignoring rare terms that appear only one time in the target texts, Table 2B recounts the number of frequent nontrivial terms. The number drops by half to 10,000+ for the four supervised methods, while only drops by 8% for the unsupervised TopWORDS. We note that the term dictionaries reported by the five different methods do not match well with each other:  $D_c^{LTP}$  and  $D_c^{THU}$  achieve the largest overlap ratio of 60%–70% for nontrivial terms and 70%–80% for frequent nontrivial terms, and the overlap ratio of all the other pairs varies from 20%–70%. These facts reflect the critical challenges in term discovery from domain-specific Chinese texts.

Combining the five term dictionaries, we obtain a joint term dictionary of 80,000+ distinct terms:

$$D_c = D_c^{JB} \cup D_c^{SP} \cup D_c^{LTP} \cup D_c^{THU} \cup D_c^{TW}.$$

We identified 20,000+ technical terms with clear medical meanings (including 14,300+ symptoms, 3,600+ body parts, 2,000+ disease names, 900+ lab

<sup>1</sup>For the technical terms: (1) we asked two TCM experts to label the discovered terms independently, if both of them agreed a discovered term to be technical, we labeled it as a technical term; (2) for most technical terms, the two experts gave the same label, and for the a few technical terms received different type labels, we asked two experts to discuss with each other for a second time and reported their consensus.

tests and 400 + medical treatments<sup>1</sup>), 60,600 + background terms (*i.e.*, correct words and phrases with no medical meanings), and 5,700 + suspicious terms whose semantic meanings cannot be easily determined. Table 1 C lists the top 30 terms for each of the 6 term categories in  $D_c$ . Table 2C shows the contribution of different methods to discovery of technical terms, background terms and suspicious terms, respectively. The results suggest that the supervised methods indeed missed a lot of meaningful technical terms in this study, while the unsupervised TopWORDS discovers 13,047 (61%) technical terms missed by other methods. Figure 3 shows the length distribution and type distribution of the discovered technical terms in  $D_c$  by different methods. From these figures, we can see that TopWORDS tends to report more longer words than the supervised methods, and contributes most to the discovery of technical terms. A term-feature map  $M_c$  for these discovered technical terms is established to achieve term standardization in a similar way to establish  $M_b$ .

The variation on term discovery naturally leads to variation on text segmentation. Table 2D and 2E

compare the performance of different methods on text segmentation based on the term boundary profile  $P_c = (V_c^{JB}, V_c^{SP}, V_c^{LTP}, V_c^{THU}, V_c^{TW})$ . We propose two different criteria for the comparison of two methods: the less rigorous criterion based on *segmentation sites*, and the more rigorous criterion based on *segmented terms*. Let  $V$  and  $V'$  be the term segmentation vectors of the two methods to be compared. The *segmentation site criterion* simply counts the number of common sites segmented by both methods, *i.e.*,  $\#\{i: V_i = V'_i = 1\}$ ; the *segmented term criterion*, however, counts the number of common terms segmented by both methods, *i.e.*,  $\#\{i: V_i = V'_i \in \{1, 2\}, \text{ and } \exists t > 0, \text{ s.t.}, V_{i+t} = V'_{i+t} \in \{1, 2\}, \text{ and } V_{i+s} = V'_{i+s} = 0 \text{ for } 0 < s < t\}$ . The degree of agreement of the five tested methods varies between 30% to 90% under the segmentation site criterion, and drops to 20%–85% under the more rigorous segmented term criterion. The supervised methods tend to segment the target texts into smaller pieces, while the unsupervised TopWORDS tends to cut the target texts with a larger granularity.

Because many technical terms are missed by each of

**Table 2 Comparison of term discovery and text segmentation of free texts by different methods**

(A) Nontrivial terms discovered by different methods from the unstructured symptoms field

	Discovered words	Overlap with Jieba	Overlap with SP	Overlap with LTP	Overlap with THULAC	Overlap with TopWORDS	Overlap with segment list
Jieba	23,989	23,989 (100%)	13,906 (68%)	13,688 (57%)	13,270 (55%)	9,864 (41%)	2,765 (12%)
SP	26,358	13,906 (53%)	26,358 (100%)	14,815 (56%)	14,217 (54%)	10,722 (41%)	4,899 (19%)
LTP	28,619	13,688 (48%)	14,818 (52%)	28,619 (100%)	20,137 (70%)	10,923 (38%)	4,142 (14%)
THULAC	30,254	13,270 (44%)	14,217 (47%)	20,137 (67%)	30,254 (100%)	12,088 (40%)	4,096 (14%)
TopWORDS	47,248	9,864 (21%)	10,722 (23%)	10,923 (23%)	12,088 (26%)	47,248 (100%)	17,365 (37%)

(B) Frequent nontrivial terms discovered by different methods from the unstructured symptoms field

	Discovered words	Overlap with Jieba	Overlap with SP	Overlap with LTP	Overlap with THULAC	Overlap with TopWORDS	Overlap with segment list
Jieba	11,412	11,412 (100%)	7,419 (65%)	7,025 (62%)	6,907 (61%)	8,062 (71%)	2,061 (18%)
SP	11,225	7,419 (66%)	11,225 (100%)	7,442 (66%)	7,261 (65%)	8,275 (74%)	2,699 (24%)
LTP	11,590	7,025 (61%)	7,442 (64%)	11,590 (100%)	9,184 (79%)	8,050 (69%)	2,456 (21%)
THULAC	12,298	6,907 (56%)	7,261 (59%)	9,184 (75%)	12,298 (100%)	8,343 (68%)	2,484 (20%)
TopWORDS	43,300	8,062 (19%)	8,275 (19%)	8,050 (19%)	8,343 (19%)	43,300 (100%)	16,593 (38%)

(C) Contribution of different methods to term discovery

	Total number	Contribution by different methods				
		Jieba	SP	LTP	THULAC	TopWORDS
Technical terms	21,454	3,500 (16%)	5,419 (25%)	5,340 (25%)	6,018 (28%)	18,844 (88%)
Background terms	60,623	21,037 (35%)	19,858 (33%)	23,485 (39%)	24,720 (41%)	27,929 (46%)
Suspicious terms	5,755	968 (17%)	2,510 (44%)	1,420 (25%)	1,121 (19%)	2,564 (45%)
Frequent technical terms	15,513	2,291 (15%)	2,767 (18%)	2,833 (18%)	3,069 (20%)	15,209 (98%)
Frequent background terms	35,004	9,871 (28%)	8,963 (26%)	9,642 (28%)	10,069 (29%)	27,623 (79%)
Frequent suspicious terms	2,940	379 (13%)	622 (21%)	407 (14%)	406 (14%)	2,554 (87%)

(D) Comparison of segmentation sites by different methods

(continued)

	Segmentation sites	Overlap with Jieba	Overlap with SP	Overlap with LTP	Overlap with THULAC	Overlap with TopWORDS
Jieba	469,381	469,381 (100%)	382,513 (81%)	396,660 (85%)	393,300 (84%)	162,473 (35%)
SP	476,058	382,513 (80%)	476,058 (100%)	418,836 (88%)	408,289 (86%)	160,548 (34%)
LTP	525,648	396,660 (75%)	418,836 (80%)	525,648 (100%)	476,232 (91%)	158,298 (30%)
THULAC	525,308	393,300 (75%)	408,289 (78%)	476,232 (91%)	525,308 (100%)	160,094 (30%)
TopWORDS	185,869	162,473 (87%)	160,548 (86%)	158,298 (85%)	160,094 (86%)	185,869 (100%)

(E) Comparison of segmented words by different methods

(continued)

	Segmented words	Overlap with Jieba	Overlap with SP	Overlap with LTP	Overlap with THULAC	Overlap with TopWORDS
Jieba	709,385	709,385 (100%)	484,564 (68%)	479,641 (68%)	475,637 (67%)	201,698 (28%)
SP	716,062	484,564 (68%)	716,062 (100%)	533,394 (74%)	512,552 (72%)	201,159 (28%)
LTP	765,652	479,641 (63%)	533,394 (70%)	765,652 (100%)	645,135 (84%)	184,021 (24%)
THULAC	765,312	475,637 (62%)	512,552 (67%)	645,135 (84%)	765,312 (100%)	185,705 (24%)
TopWORDS	425,873	201,698 (47%)	201,159 (47%)	184,021 (43%)	185,705 (44%)	425,873 (100%)

these tested methods, it's risky to proceed the downstream analysis based on the segmented texts by any of these methods alone. To get rid of this dilemma, we feed to TopWORDS the joint term dictionary  $D_c$  as the preload vocabulary, and refit the model on the free texts from the segment list  $S_c$  to do text segmentation. We choose TopWORDS as the segmentation tool because the segmentation results from it enjoy a proper granularity for semantic understanding of the free texts in the archive.

Totally, 8,000 + features are generated for the 20,000 + technical terms discovered, resulting in a feature code-book  $F_c$ . The transformation rules from terms in  $D_c$  to features in  $F_c$  are summarized in a term-feature map  $M_c$ . Mapping the technical terms in the segmented texts to their standard feature codes based on  $M_c$  with all the background terms ignored, we can transform the free texts in the unstructured symptoms field into a structured feature table  $T_c$  of binary features (with 1 for presence of a feature, and 0 for absence).

### Generating a united feature table via data integration

The structured feature tables  $T_a$ ,  $T_b$  and  $T_c$  generated from the structured, semi-structured and unstructured data fields of the archive can be further integrated into a united feature table  $T = (T_a, T_b, T_c)$  as the final output of the data processing procedure. Some features may be shared by  $T_a$ ,  $T_b$  and  $T_c$ . We combined information about these shared features via data integration.

Totally, 14,000 + distinct features are involved in the united feature table for the 26,000 + visits. And, 26,000 + transformation rules for technical terms are created to establish the table. Detailed contents about the united feature codebooks  $F = F_a \cup F_b \cup F_c$ , the united

term dictionary  $D = D_b \cup D_c$ , the united term-feature map  $M = M_b \cup M_c$ , and the united structured feature table  $T$  can be found in the website of "Zhou Archive for TCM Study".

## STATISTICAL LEARNING OF THE STRUCTURED FEATURE TABLE

Based on the structured feature table obtained, a series of statistical analysis can be implemented to learn potential principles of TCM clinical practice from a data-driven perspective. Considering that the missing rate of some data fields is very high, to avoid the potential influence of these missing values, in this study we only select the technical terms from the following 7 data fields whose missing rate in first-visit records are less than 30%: Symptoms, Tongue Picture, Pulse Type, Disease, Disease Category, TCM Pathogenesis and Herbs in TCM Prescription. Totally, 7,743 features are involved in these selected data fields, among which 1,926 are rare features whose frequency in the first-visit records  $\leq 1$ . We ignored these rare features in the downstream analysis, and only focused on the 5,817 frequent features.

### Correlation analysis

Our first effort is a correlation analysis to capture the overall correlation structures of all the 5,817 selected cross-category features. A  $5,817 \times 5,817$  correlation matrix is obtained and most correlation coefficients in the matrix fall into  $[-0.1, 0.1]$  indicating relatively weak correlation. Figure 4 shows the correlation heat map of a few highly-correlated features which correlate with some other features with a correlation coefficient beyond  $(-0.5, 0.5)$ .

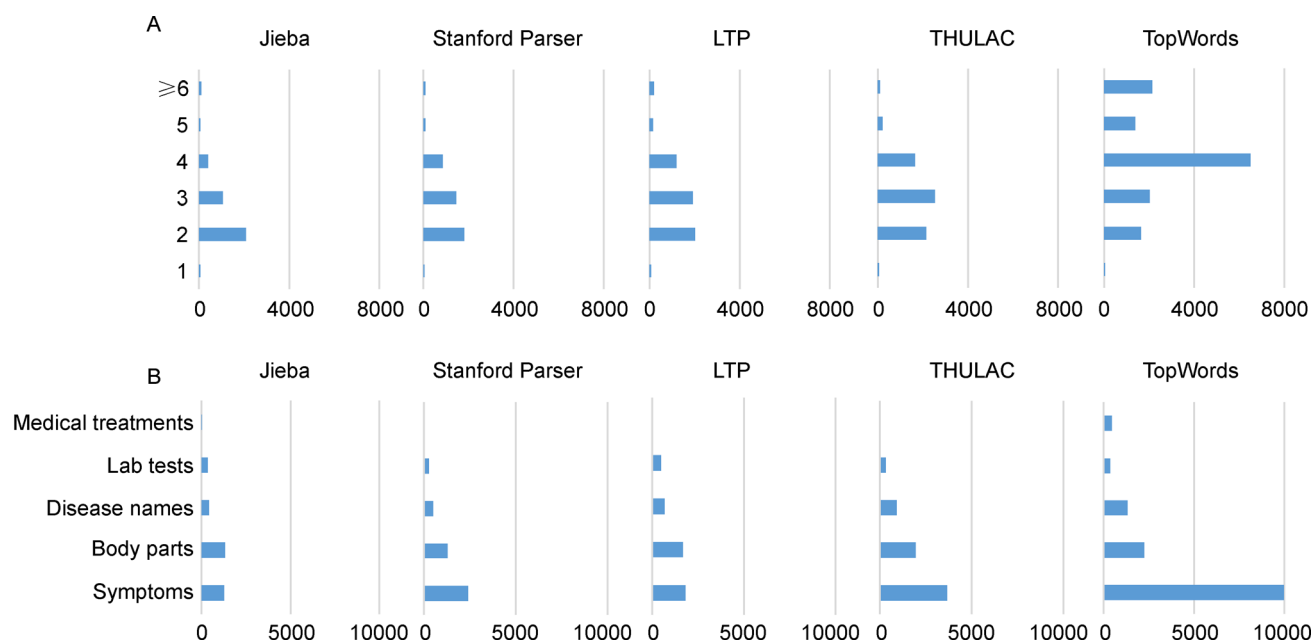
From the heat map, we can observe a few blocks of highly correlated features. For example, the largest feature block highlighted in a black box reveals that disease category “cancers” are closely related to TCM pathogenesis “toxic head” and “phlegm stasis”, and a group of herbs “*Andropogon squarrosus* radix”, “*Glehnia littoralis* radix”, “*Buttercup* root”, “*Pseudostellariae* radix”, “*Ophiopogon* radix”, “*Appendiculate cremastra pseudobulb*”, “*Herba euphorbiae helioscopiae*”, “*Agrimony*”, “*Hedyotis*”, “*Barbated skullcup herb*”, “*Herba celiptae*”, “*Ligustri lucidi fructus*”. The smaller feature block next to the largest one discovers that a few features on Pulse Type are closely related. The other feature block located at the left-bottom corner reveal a group of herbs (*i.e.*, “*Lignum phetimiae*”, “orientvine stem”, “largeleaf gentian root”, “*Preparemonkshd moter* root”, “*Aconiti preparata*”, “*Asarum sieboldii*”) closely related to disease category “rheumatism”. These discoveries reveal meaningful TCM knowledge.

### Enrichment analysis

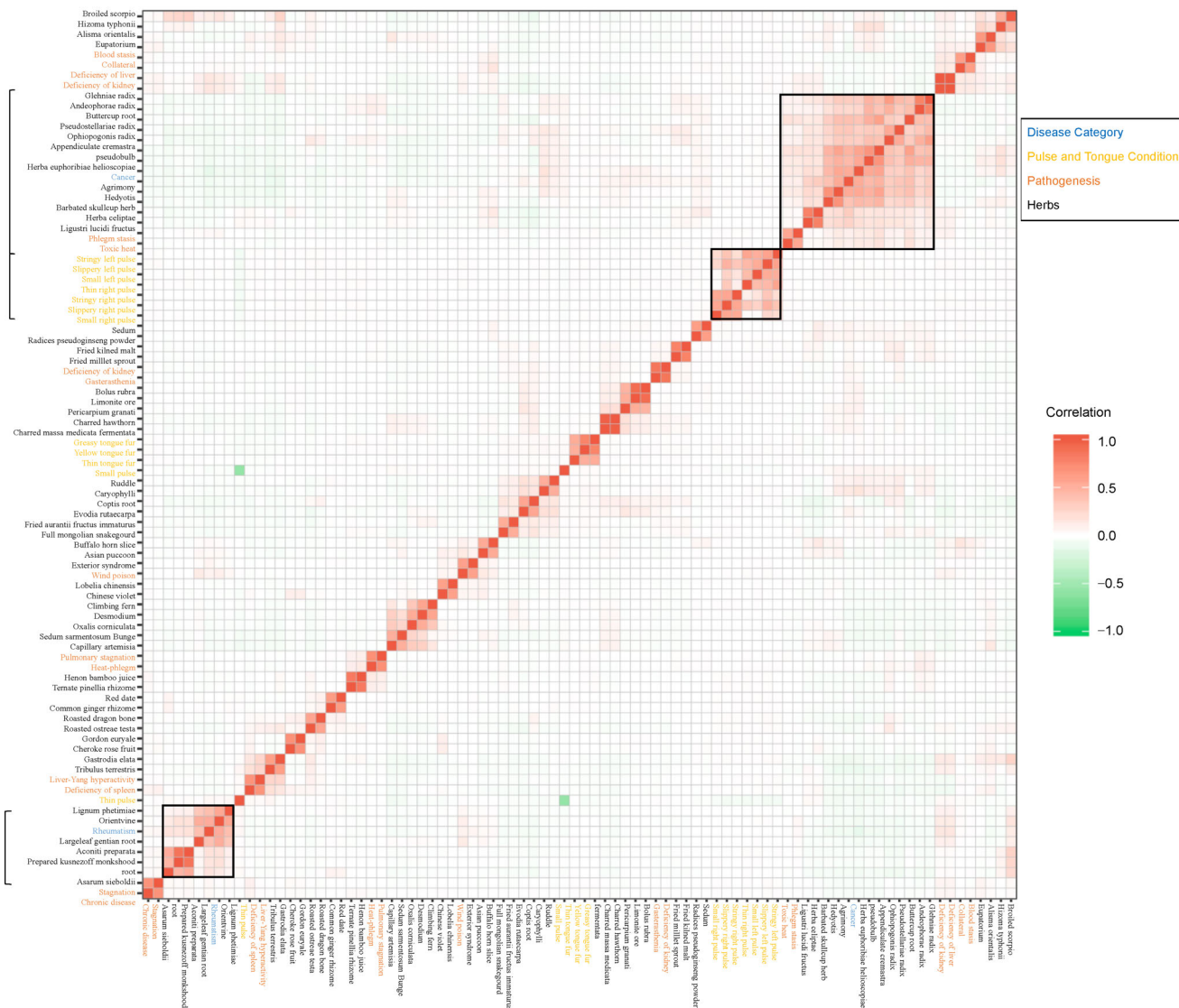
To further investigate how TCM concepts such as TCM diseases, TCM pathogenesis and TCM therapies connect to each other and other features such as diseases, symptoms and herbs, we did the enrichment analysis below. For simplicity, let’s take “stiffness of stomach”, the most frequent TCM disease in the archive, as an example. First, we selected from the archive all first-visit records of which the feature “stiffness of stomach” takes 1. We denote the subpopulation of selected records as  $P_1$ ,

the subpopulation of other first-visit records as  $P_0$ . Next, we identified the top 5 diseases, symptoms, TCM pathogenesis, TCM pathogenesis and herbs in the selected records in  $P_1$  with the highest relative frequency. Third, for each of the selected feature, we calculated its odds ratio between  $P_1$  and  $P_0$  as its enrichment measure with respect to feature “stiffness of stomach”. At last, we plotted the relative frequency and the enrichment measure for all feature selected for “stiffness of stomach” in a bar plot as showed in Figure 5A. Such an *enrichment plot* demonstrates rich information about TCM disease “stiffness of stomach”: (i) it is associated with chronic gastritis, chronic superficial dermatitis, chronic atrophic antralgastritis, headache and astriction, (ii) symptom gastric distention is major signature of it, (iii) “liver-stomach disharmony”, “dampness and heat resistance” and “stomach weak energy stagnation” are the major TCM pathogenesis behind it, and (iv) “processed pinellia preparata”, “*Cyperus rhizoma*”, “*Perilla caulis*”, “*Coptidis* root” and “*Magnolia bark*” are the primary herbs to treat it. These messages help us understand the basic properties of the feature efficiently from multiple angles.

Figure 5 displays the enrichment plots for a few most frequent TCM diseases, TCM pathogenesis and TCM therapies in the archive. We can read many insightful messages from these figures. For example, TCM pathogenesis “dampness-heat” is highly associated with liver-related diseases, and takes “impairment of liver and spleen” and “*Phellodendri chinensis cortex*” as the signature symptom and herb respectively; TCM therapy



**Figure 3. Length and type distribution of technical terms in  $D_c$  discovered by different methods.** (A) Length distribution. (B) Type distribution



**Figure 4.** Correlation heat map of highly correlated features based on the first-visit records.

“regulating and harmonizing the liver and spleen” is a regular treatment for liver-related diseases and TCM pathogenesis, and takes “barbary wolfberry fruit”, “pseudostellariae radix”, “deep-fried atractylodis macrocephalae rhizoma”, “salviae miltiorrhizae” and “paeoniae radix rubra” as the primary components of prescription.

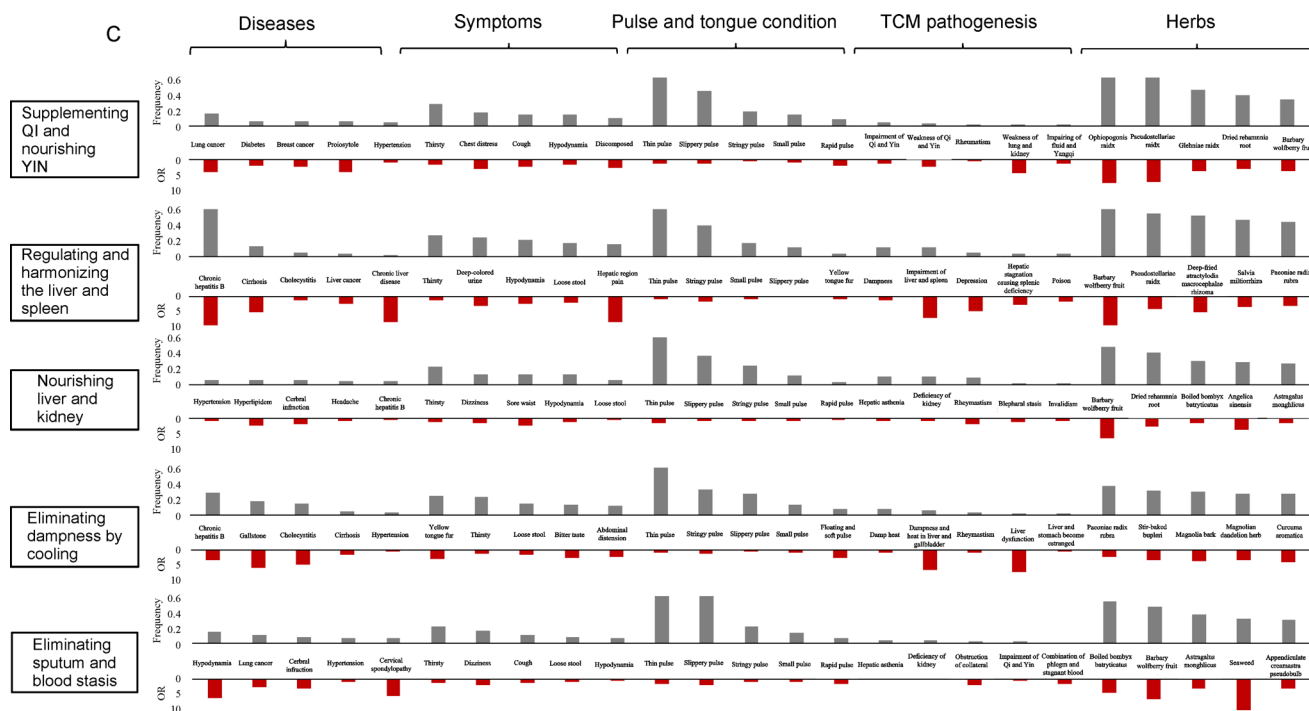
## Embedding analysis

Considering that correlation and enrichment analyses basically utilize the data information in a pairwise fashion, they may miss signals reflecting high-order structures in the data. In this section, we analyzed the structured feature tables obtained from the Zhou Archive from an alternative perspective via embedding methods

[32–34]. Different from previous correlation and enrichment analyses, embedding analysis considers co-occurrence patterns of different features globally, and embeds features with no geometric meanings (e.g., symptoms and herbs) into a linear space with geometric interpretation.

Treating each feature as a “word” and each record as a “document”, we can naturally apply approaches designed for *word embedding* to the TCM data. Here, we selected the matrix factorization approach [32] as the primary tool to achieve feature-level embedding, and applied it to the  $4,776 \times 4,776$  co-occurrence matrix  $C$  of the frequent features, where  $C_{ij}$  counts the number of first-visit records that contain both feature  $i$  and feature  $j$ , to embed the frequent features into a 300-dimensional linear space. The detailed embedding vectors of the features can be found in





**Figure 5.** Enrichment analysis for the top 5 TCM diseases, TCM pathogenesis and TCM therapies. (A) Enrichment analysis for the top 5 TCM diseases. (B) Enrichment analysis for the top 5 TCM pathogenesis (C) Enrichment analysis for top 5 TCM therapies.

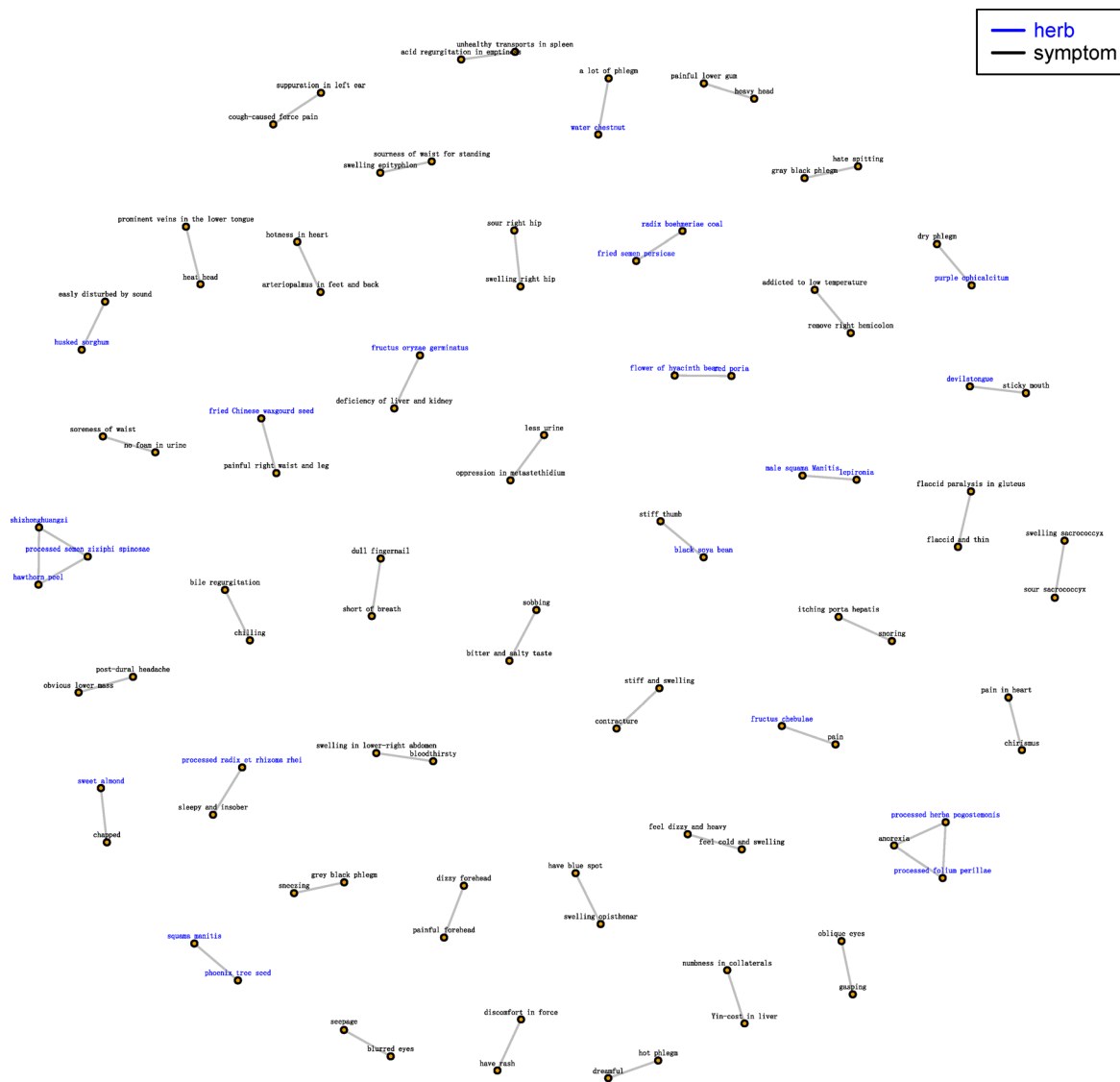
the website of “Zhou Archive for TCM Study”. Features that stay close to each other in the embedding space tend to associate closely or share similar functions. Geometric structure of these embedding vectors can be visualized in a 2-dimensional space by techniques such as *multi-dimensional scaling* (MSD) [34]. Figure 6 shows the MSD plot of the 50 feature pairs with the shortest within-pair distance in the embedding space, most of which precisely reflect TCM knowledge. For example, the two symptoms in pair {painful forehead, dizzy forehead} are complication that often happen concurrently, the two herbs in pair {processed herba pogostemonis, processed folium perillae} have similar function in expelling cold and vomiting, and the symptom-herb pair {chapped, sweet almond} corresponds to a well-known treatment to chapped and irritated skin in TCM.

Beyond the feature-level embedding, we can also embed records into a linear space in a similar way. Representing each record by a vector of 4,776 binary variables with 1s and 0s standing for the presence and absence of features in the record, we used t-Distributed Stochastic Neighbor Embedding (t-SNE) [33] to embed the high-dimensional records into a 2-dimensional representation space. Unlike the linear dimension reduction technique Principal Component Analysis (PCA) by maximizing variance to preserve large pairwise distances which fails in non-linear structure cases, t-SNE tries to

retain the local structures while preserving almost the same topology by embedding the original high dimensional space with a Student t-distribution. Figure 7 demonstrates the results from t-SNE, where each point corresponds to a record with the color stands for the disease category of the record. Among the 16 distinct diseases categories in the archive, the 5 major categories, *i.e.*, Cancers, Digestive Diseases (DD), Infectious Diseases (InD), Neurological Diseases (ND) and Respiratory Diseases (RD), together with miscellaneous Diseases (MiD) contribute ~75% of the records. Interestingly, points associated with the 5 major categories cluster well in the embedding space, with the MiD-related points spread out everywhere. These phenomena reflect heterogeneity of TCM practice among different disease categories and are consistent to the definition of MiD.

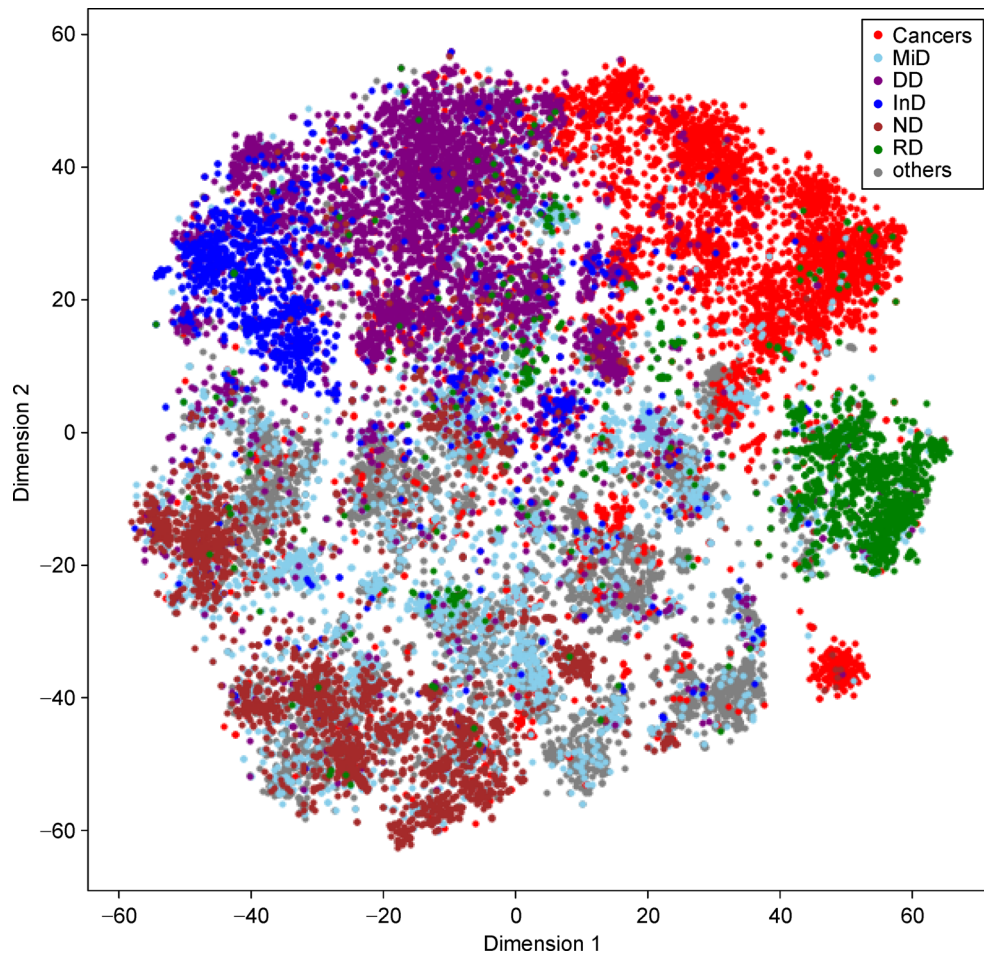
### Association pattern discovery

Next, we try to discover association patterns of the selected features from structured feature table. As all features in the feature table is binary, the data structure perfectly fit the classic *Market Basket Analysis* (MBA) problem [35] in machine learning, which aims to discover items that tend to purchased together from a collection of baskets purchased by customers to a supermarket. *Association Rule Mining* (ARM) [35,36] is the classic



Recently, Refs. [37,38] reformulated the MBA problem into a statistical model selection problem and proposed a novel solution to this classic problem from the statistical point of view. Assuming that each basket is composed of a collection of item modules (called *themes*) randomly selected by the customer with different selection probabilities and different baskets are generated independently

With the support of EMRs in the Zhou Archive which contains both symptoms and prescriptions, in this study,



**Figure 7.** Embedding EMRs in the Zhou Archive into 2-dimensional space by t-SNE

**Table 3** The top 60 cross-category themes discovered by CTDM

Symptoms	Herbs
Dry mouth	Glehniae radix, andeophorae radix
	Dried rhizome of rehmanni, salvia chinensis
	Dendrobium
	Radix trichosanthis, asparagus cochinchinensis, ophiopogonis radix
	Radix trichosanthis, rhizoma anemarrhenae
	Calcined oyster, calcined fossil fragment
	Asparagus cochinchinensis, ophiopogonis radix
	Figwort root, flos chrysanthemi indicis
	Hiraut shiny bugleweed herb, alismatis
Poor sleep	Tuber fleeceflower stem
	Cooked date seed
	Asparagus cochinchinensis, lilium davidii, cooked date seed
	Cortex albiziae
	Asparagus cochinchinensis, cooked date seed
	Asparagus cochinchinensis, lilium davidii, cooked date seed, cortex albiziae

(Continued)

Symptoms	Herbs
Stomach distension	Processed rhizoma, caulis perillae Coptidis Fried fructus aurantii immaturus, rhizoma pinellinae praeparata Dried orange peel, rhizoma pinellinae praeparata Dried orange peel, immature tangerine peel
Debility	Eclipta alba, processed glossy privet fruit Fried atractylodes macrocephala koidz, poria cocos, codonopsis, radix glycyrrhizae Red paeonia, processed rhizoma, vinegar-baked bupleurum root Fried atractylodes macrocephala koidz, poria cocos, radix glycyrrhizae, pseudostellariae radix Asparagus cochinchinensis, ophiopogonis radix
Belching	Rhizoma pinellinae praeparata Coptidis Fructus amomi
Gastralgia	Processed rhizoma, caulis perillae Rhizoma pinellinae praeparata Fructus amomi, costus root
Dizziness	Tribulus terrestris, gastrodiae, ligusticum wallichii Tribulusterrestris, chrysanthemum, gastrodiae Gastrodiae, ligusticum wallichii
Sensation of chill	Radix glycyrrhizae, processed cassia twig Parched white peony root, processed cassia twig Cinnamon
Palpitation	Salvia miltiorrhiza, ligusticum wallichii Salvia miltiorrhiza
Headache	Tribulusterrestris, gastrodiae, ligusticum wallichii Ligusticum wallichii
Poor appetite	Fried atractylodes macrocephala koidz, poria cocos, radix glycyrrhizae Pseudostellariae radix, coloured malt, fried millet sprout
Yellowish complexion	Astragali radix Chinese angelica
Cough	Glehniae radix, ophiopogonis radix Glehniae radix
Dry stool	Fructus trichosanthis Roasted fructus aurantii immaturus, fructus trichosanthis
Vertigo	Barbary wolfberry fruit, gastrodiae
Deep-colored urine	Radix sophorae flavescentis
Bitter in mouth	Fructus evodiae, coptidis
Abdominal distension	Dried orange peel, immature tangerine peel
Feel agitated	Asparagus cochinchinensis, lilium davidii
Borborygmus	Fructus evodiae, coptidis
Loose stool	Fried atractylodes macrocephala koidz, codonopsis
Cough, oppression in chest	Rhizoma pinellinae praeparata
Oppression in chest	Red paeonia, processed rhizoma, vinegar-baked bupleurum root
oppression in chest, palpitation	Salvia miltiorrhiza
Nausea, vomiting	Rhizoma pinellinae praeparata

we generalize this idea to learn association patterns between a module of symptoms and a module of herbs. By treating symptom-related features and herbs in the prescriptions as “items” and each EMR as a “basket” of these items, we obtained 23,000 + effective “baskets” from the first-visit records in the archive. The original TDM can discover themes of all items in the baskets from one single category. In this study, however, we are more interested in *cross-category themes* containing both symptoms and herbs, which connect a module of symptoms to a module of herbs and provide information on how TCM treatment is determined based on the observed symptoms. To fit this special request, we modified the original TDM approach to a variant version by adding some filters to label items from different categories which rules out all single-category themes via a pre-screening of themes in the initial theme dictionary. After removing those redundant single-category association rules in the initial theme dictionary, it largely reduces the number of the partitions for all baskets. We refer to this variant version of the original TDM approach as to the *Cross-category TDM* approach, which is abbreviated to CTDM. Compared to the original TDM approach, the CTDM approach enjoys a better computational efficiency as many single-category themes are excluded from the model priori.

Please note that we meant to include all diagnosis-related features here to link symptoms to herbs directly. And, we only kept the first-visit records here, because the longitudinal records of the same patient are often highly correlated with each other (the physical condition of a patient typically does not change dramatically within a couple of months, leading to similar symptoms, diagnoses and treatments), and may seriously violate the assumption of independent samples behind TDM. We also removed baskets containing more than 30 items which may largely slow down the procedure. Totally, 5,175 effective items survived this item-basket screening procedure, resulting in a collection of baskets with 20 items on average.

Same as the TDM approach, the CTDM approach has two control parameters: the *minimum theme frequency parameter*  $\tau_P$  and the *maximum theme length parameter*  $\tau_L$ . In this study, we set  $\tau_L = 6$  and  $\tau_P = 0.001$ , and discovered ~1,000 cross-category themes from the archive. Table 3 shows the top 60 cross-category themes discovered by the CTDM approach, each of which connects a module of symptoms to a module of herbs, revealing important insights of TCM treatment. For example, the connections between herb modules {asparagus cochinchinensis, lilium davidii} and symptom modules {poor sleep} and {feel agitated}, the connection between symptom module {dry mouth} and herb modules {figwort root, flos chrysanthemi indici}, the connections between herb modules {tribulus terrestris, gastrodiae,

ligusticum wallichii} and symptom modules {dizziness} and {headache}, the connection between symptom modules {oppression in chest, palpitation} and herb module {salvia miltiorrhiza} all precisely reflect important principles in TCM practice.

Please note that we clustered these top themes based on the symptom module and rearranged their location in the table to deliver information more efficiently. The complete list of discovered themes can be found in the website of “Zhou Archive for TCM Study”.

## CONCLUSION AND DISCUSSIONS

In this study, we introduce the Zhou Archive, a large-scale database of expert-specific EMRs containing comprehensive information about 73,000 + visits to one TCM doctor by 26,000 + distinct patients over 35 years from 1980 to 2015. Processing the text data in the archive via a series of data processing steps with the help of multiple popular NLP tools for Chinese texts, we transformed the semi-structure EMRs in the archive to a well-structured feature table. A series of statistical analyses are implemented for the structured feature table obtained to learn principles of TCM clinical practice from the archive. Results from these analyses reveal insights to understand TCM from a data-driven perspective. Besides the statistical analysis demonstrated in this paper, many other methods and new tools can be applied or developed to dig deeper into this archive. We hope the data processing and analysis framework proposed in this paper can motivate other studies for understanding TCM based on large-scale EMR datasets.

## SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at <https://doi.org/10.1007/s40484-019-0173-x>.

## ACKNOWLEDGEMENT

We thank the Zhou Zhongying’s Studio at Nanjing University of Chinese Medicine for the great efforts on collecting, managing and sharing this valuable archive. We also thank Miss Bing Liang, Mr. Qiuyu Liang and Miss Che Wang for their efforts on data preparation and preprocessing.

This work was partially supported by the National Natural Science Foundation of China (Nos. 11771242 & 11401338), the Tsinghua University Initiative Scientific Research Program and Supporting Grant to the Zhou Zhongying’s Studio 201159 by the State Administration of TCM of China.

## COMPLIANCE WITH ETHICS GUIDELINES

The authors Yang Yang, Qi Li, Zhaoyang Liu, Fang Ye and Ke Deng declare that they have no conflict of interests.

All procedures were in accordance with the ethical standards of the

institution or practice at which the studies were conducted, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

## APPENDIX

### Introduction to the five NLP tools involved

**Jieba** is an open-source software developed by Sun Junyi in 2012. The software uses a variant of *maximum matching algorithm* and *dynamic programming* to achieve word segmentation, and use a *Hidden Markov Model* to achieve named entity recognition. The method is equipped with a preloaded vocabulary of more than 20,000 words, and trained with manually segmented and labelled news articles from *People's Daily* and some Chinese novels segmented by ICTCLAS. Here, we use the “accurate mode” of Jieba version 0.38 to do all the analysis, and denote the reported dictionary as  $D^{JB}$ .

**Stanford Parser** is a tool developed by the Stanford Natural Language Processing Group in 2003, which is a multi-language parser that can be used in English, Chinese, German, etc. Trained with the Penn Chinese Treebank, Stanford Parser can work out the grammatical structure of Chinese sentences on top of word segmentation. Here, we use the Stanford Parser version 3.9.1 to do all the analysis, and denote the reported dictionary as  $D^{SP}$ .

**LTP** is another open-source platform developed by the Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology in 2007. It uses *forward maximum match* to merge the information of a preloaded vocabulary into the statistic model, and is equipped the online learning technique for faster computing. Here, we use the LTP version 3.4.0 to do all the analysis and denote the reported dictionary as  $D^{LTP}$ .

**THULAC** is a tool developed by the Natural Language Processing Group at the Department of Computer Science and Technology in Tsinghua University in 2016. It achieves word segmentation based on the *maximum entropy approach* [12]. The statistical model behind is trained with manually segmented and labelled news articles from *People's Daily* and other sources, which contain a total amount of 58 million Chinese characters. Here, we use THULAC python version v1\_2 to do all the analysis, and denote the reported dictionary as  $D^{THU}$ .

**TopWORDS** is a tool developed by Ke Deng and Jun S. Liu in 2016. Different from the above supervised methods which emphasize *precise word segmentation* under the guidance of a preloaded vocabulary and high-quality training corpus, TopWORDS pays more attention on *efficient new word discovery* when the preloaded vocabulary and the training corpus do not fit the target texts well. Starting with an over-complete initial dictionary generated by enumerating all frequent strings in

the target texts, and pruning it into a much smaller final dictionary via statistical model selection, TopWORDS can effectively discover previously unknown words and phrases that appear in the target texts more than 3 times when no preloaded vocabulary and proper training corpus are available (available preloaded vocabulary and training corpus will improve the performance of TopWORDS). TopWORDS has two control parameters: the minimal word frequency  $\tau_P$  and the maximum word length  $\tau_L$ . We specify  $\tau_P = 3$  and  $\tau_L = 8$  in this study.

## REFERENCES

1. Liu, W. H. (2017) TCM acupuncture-moxibustion: contributing to human health. *World J. Acupunct. Moxibustion*, 27, 1
2. Ahn, A. C., Bennani, T., Freeman, R., Hamdy, O. and Kaptchuk, T. J. (2007) Two styles of acupuncture for treating painful diabetic neuropathy—a pilot randomised control trial. *Acupunct. Med.*, 25, 11–17
3. Liu, Z., Sun, F., Zhu, M. and Wang, X. (2004) Effect of acupuncture on insulin resistance in non-insulin dependent diabetes mellitus. *J. Acupunct. Tuina Sci.*, 2, 8–11
4. Li, S. and Zhang, B. (2013) Traditional Chinese medicine network pharmacology: theory, methodology and application. *Chin. J. Nat. Med.*, 11, 110–120
5. Zhang, B., Wang, X. and Li, S. (2013) An integrative platform of TCM network pharmacology and its application on a herbal formula, Qing-Luo-Yin. *Evid. Based Complement. Alternat. Med.*, 2013, 456747
6. Li, S., Zhang, B. and Zhang, N. (2011) Network target for screening synergistic drug combinations with application to traditional Chinese medicine. *BMC Syst. Biol.*, 5, S10
7. Lam, W., Bussom, S., Guan, F., Jiang, Z., Zhang, W., Gullen, E. A., Liu, S. H. and Cheng, Y. C. (2010) The four-herb Chinese medicine PHY906 reduces chemotherapy-induced gastrointestinal toxicity. *Sci. Transl. Med.*, 2, 45ra59
8. Xiang, Y. Z., Shang, H. C., Gao, X. M. and Zhang, B. L. (2008) A comparison of the ancient use of ginseng in traditional Chinese medicine with modern pharmacological experiments and clinical trials. *Phytother. Res.*, 22, 851–858
9. Jian, J. and Wu, Z. (2004) Influences of traditional Chinese medicine on non-specific immunity of Jian Carp (*Cyprinus carpio* var. Jian). *Fish Shellfish Immunol.*, 16, 185–191
10. Bick, R. J., Poindexter, B. J., Sweney, R. R. and Dasgupta, A. (2002) Effects of Chan Su, a traditional Chinese medicine, on the calcium transients of isolated cardiomyocytes: cardiotoxicity due to more than Na, K-ATPase blocking. *Life Sci.*, 72, 699–709
11. Iwasaki, K., Satoh-Nakagawa, T., Maruyama, M., Monma, Y., Nemoto, M., Tomita, N., Tanji, H., Fujiwara, H., Seki, T., Fujii, M., et al. (2005) A randomized, observer-blind, controlled trial of the traditional Chinese medicine Yi-Gan San for improvement of behavioral and psychological symptoms and activities of daily living in dementia patients. *J. Clin. Psychiatry*, 66, 248–252

12. Deng, K., Liu, D., Gao, S. and Geng, Z. (2005) Structural learning of graphical models and its applications to traditional Chinese medicine. *Lect. Notes Comput. Sci.*, 3614, 362–367
13. Feng, Y., Wu, Z., Zhou, X., Zhou, Z. and Fan, W. (2006) Knowledge discovery in traditional Chinese medicine: state of the art and perspectives. *Artif. Intell. Med.*, 38, 219–236
14. Yang, H., Chen, J., Tang, S., Li, Z., Zhen, Y., Huang, L. and Yi, J. (2009) New drug R&D of traditional Chinese medicine: role of data mining approaches. *J. Biol. Syst.*, 17, 329–347
15. Wang, Q. and Zhu, Y. (2009) Epidemiological investigation of constitutional types of Chinese medicine in general population: based on 21,948 epidemiological investigation data of nine provinces in China. *Zhonghua Zhongyiyao Zazhi* (in Chinese), 24, 7–12
16. Xue, R., Fang, Z., Zhang, M., Yi, Z., Wen, C. and Shi, T. (2013) TCMID: traditional Chinese Medicine integrative database for herb molecular mechanism analysis. *Nucleic Acids Res.*, 41, D1089–D1095
17. Liu, B., Zhou, X., Wang, Y., Hu, J., He, L., Zhang, R., Chen, S. and Guo, Y. (2012) Data processing and analysis in real-world traditional Chinese medicine clinical data: challenges and approaches. *Stat. Med.*, 31, 653–660
18. Wang, X., Qu, H., Liu, P. and Cheng, Y. (2004) A self-learning expert system for diagnosis in traditional Chinese medicine. *Expert Syst. Appl.*, 26, 557–566
19. Yu, S., Ma, Y., Gronsbell, J., Cai, T., Ananthakrishnan, A. N., Gainer, V. S., Churchill, S. E., Szolovits, P., Murphy, S. N., Kohane, I. S., *et al.* (2018) Enabling phenotypic big data with PheNorm. *J. Am. Med. Inform. Assoc.*, 25, 54–60
20. Roden, D. M., Pulley, J. M., Basford, M. A., Bernard, G. R., Clayton, E. W., Balser, J. R. and Masys, D. R. (2008) Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.*, 84, 362–369
21. Blair, D. R., Lyttle, C. S., Mortensen, J. M., Bearden, C. F., Jensen, A. B., Khiabani, H., Melamed, R., Rabadan, R., Bernstam, E. V., Brunak, S., *et al.* (2013) A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. *Cell*, 155, 70–80
22. Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S. and Sontag, D. (2017) Learning a health knowledge graph from electronic medical records. *Sci. Rep.*, 7, 5994
23. Blecker, S., Katz, S. D., Horwitz, L. I., Kuperman, G., Park, H., Gold, A. and Sontag, D. (2016) Comparison of approaches for heart failure case identification from electronic health record data. *JAMA Cardiol.*, 1, 1014–1020
24. Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., Field, J. R., Pulley, J. M., Ramirez, A. H., Bowton, E., *et al.* (2013) Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.*, 31, 1102–1110
25. Doshi-Velez, F., Ge, Y. and Kohane, I. (2014) Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, 133, e54–e63
26. Chang, P. C., Tseng, H., Dan, J. and Manning, C. D. (2009) Discriminative reordering with Chinese grammatical relations features. In: *SSST'09 Proceedings of the 3rd Workshop on Syntax and Structure in Statistical Translation*. pp. 51–59
27. Levy, R. and Manning, C. D. (2003) Is it harder to parse Chinese, or the Chinese Treebank? In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 1, 439–446
28. Che, W., Li, Z. and Liu, T. (2010) LTP: A Chinese language technology platform. In: *COLING'10 Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pp. 13–16
29. Sun, M., Chen, X., Zhang, K., Guo, Z., Ma, J. and Liu, Z. (2016) THULAC: An efficient lexical analyzer for Chinese
30. Li, Z. and Sun, M. (2009) Punctuation as implicit annotations for Chinese word segmentation. *Comput. Linguist.*, 35, 505–512
31. Deng, K., Bol, P. K., Li, K. J. and Liu, J. S. (2016) On the unsupervised analysis of domain-specific Chinese texts. *Proc. Natl. Acad. Sci. USA*, 113, 6154–6159
32. Levy, O. and Goldberg, Y. (2014) Neural word embedding as implicit matrix factorization. In: *Adv. Neural Inf. Process. Syst. Conference*
33. Maaten, L. and Hinton, G. E. (2008) Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.*, 9, 2579–2605
34. Borg, I. and Groenen, P. (1987) Modern multidimensional scaling: theory and applications. *J. Educ. Meas.*, 40, 277–280
35. Agrawal, R., Imielinski, T. and Swami, A. (1993) Mining association rules between sets of items in large databases. In: *SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pp. 207–216
36. Agrawal, R. and Srikant, R. (1994) Fast algorithms for mining association rules. In: *Readings in database systems* (3rd ed.), pp. 580–592. San Francisco: Morgan Kaufmann Publishers Inc.
37. He, P., Deng, K., Liu, Z., Liu, D., Liu, J. S. and Geng, Z. (2012) Discovering herbal functional groups of traditional Chinese medicine. *Stat. Med.*, 31, 636–642
38. Deng, K., Geng, Z. and Liu, J. S. (2014) Association pattern discovery via theme dictionary models. *J. R. Stat. Soc. B*, 76, 319–347