PaReCat: Patient Record Subcategorization for Precision Traditional Chinese Medicine

Edward W. Huang  
University of Illinois at Urbana-Champaign  
Urbana, Illinois  
ewhuang3@illinois.edu

Sheng Wang  
University of Illinois at Urbana-Champaign  
Urbana, Illinois  
swang141@illinois.edu

Runshun Zhang  
Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing, 100700, China  
runshunzhang@139.com

Baoyan Liu  
National Data Center of Traditional Chinese Medicine  
Beijing, 100700, China  
liuby@mail.cintcm.cn

Xuezhong Zhou  
Beijing Jiaotong University  
Beijing, 100044, China  
xzzhou@bjtu.edu.cn

ChengXiang Zhai  
University of Illinois at Urbana-Champaign  
Urbana, Illinois  
czhai@illinois.edu

ABSTRACT

Traditional Chinese medicine (TCM), a style of medicine widely used in China for thousands of years, can complement modern western medicine by taking personalization as the core principle of clinical practice. A fundamental task in TCM, particularly important for achieving effective precision medicine, is to subcategorize patients with a general disease into groups corresponding to variations of that disease. In this paper, we conduct the first study of the problem of subcategorizing electronic patient records in TCM. While the general problem of subcategorization can be solved using basic clustering algorithms, accommodating variations in symptoms and herb prescriptions of TCM patient records when computing patient similarity is a major technical challenge that has yet to be addressed. To tackle this problem, we propose to learn inexact matchings of both symptoms and herbs from a TCM dictionary of herb functions by using an embedding algorithm. Our hypothesis is that the prior knowledge of herb-symptom associations in the TCM dictionary can be used to discover latent relationships among comorbid symptoms and functionally similar herbs, thereby improving the quality of subcategorization. We performed extensive experiments on large-scale real-world datasets. As expected, our approach leads to more accurate matchings between patient records than baseline approaches, and thus better subcategorization results.

We also show that the proposed algorithm can be used immediately in multiple clinical applications, such as retrieving similar patients as well as discovering two special TCM cases: similar symptoms treated by different herbs and different symptoms treated by similar herbs.

CCS Concepts

- Information systems → Data analytics; Clustering; Data centers;
- Keywords
  patient record subcategorization; traditional Chinese medicine; network embedding

1. INTRODUCTION

Any single disease will manifest itself differently in different patients. This phenomenon is the result of complex interactions between a patient’s environment and numerous physiological and pathological factors, including genetic variation [14]. Consequently, patient diagnosis and treatment in real-world settings are extremely difficult due to population variation, creating demand for precision medicine.

In contrast, traditional Chinese medicine (TCM), a style of medicine widely used in China, has been leveraging personalized treatment as the core principle of clinical practice for thousands of years. Although the exact number of people who seek TCM treatment in the United States is unknown, it was estimated in 1997 that some 10,000 practitioners served more than one million patients each year [1]. TCM is often used as a complementary and alternative medicine (CAM) and is especially effective for certain diseases such as stomach ailments.

TCM doctors prescribe mixtures of herbs tailored to each patient based on a thorough assessment of his or her symptoms and physical condition; this is true even for common diseases. As a result, two patients with similar symptoms may receive completely different treatments due to the personalized nature of TCM. This manner by which TCM observes an individual’s symptom patterns is reminiscent of precision medicine techniques. Thus, it could complement
modern precision medicine, which currently relies mostly on molecular profiles [5].

Like in western medical settings, TCM doctors document symptoms that they observe from their patients. However, diseases consist of two components in TCM. The first is the “disease entity” (病), which is simply a set of symptoms. The second, unique to TCM, is the “syndrome” (证), which has no direct parallel in western medicine as the translation would suggest. Syndromes are best described as disharmonies at the core of the body. In western medicine, doctors often must resort to prescribing treatments based on symptoms alone in the event of failed differential diagnoses. On the other hand, TCM practitioners always assess and treat patients based on combinations of the two aforementioned components.

The personalized, holistic treatment of a patient in TCM is especially useful for patients with multiple, seemingly unrelated symptoms. For example, a particular patient might have facial acne, throat irritation, mouth sores, dry mouth, and abdominal pain (Figure 1). In western medicine, this patient may be referred to a skin specialist and a stomach specialist, each of whom might prescribe superficial medications to alleviate the symptoms. On the other hand, TCM doctors prescribe treatments for the patient’s symptoms (disease entities) by determining their underlying causes (syndromes).

![Figure 1: A comparison between TCM and modern western medicine.](image)

Unfortunately, this added dimension introduces an extra layer of complexity to TCM. Because two patients with the same disease might have different underlying patterns, TCM doctors will prescribe different herbs. This complication generates a necessity for subcategorizing patient records, which separates patients into groups based on their blanket diseases, and then further categorize them into smaller clusters. For example, doctors will further subcategorize the patient in Figure 1 into the “excess stomach heat” or “excess heart heat” categories.

In addition, subcategorizations can help inexperienced doctors learn pattern analysis by providing specific medical cases. Most importantly, patient record subcategorization can help doctors view all possible diagnoses for their patients, reducing the chance of misdiagnosis. Misdiaognoses across all fields not only present ongoing risks to the health and safety of patients, but also cost the United States roughly $750 billion annually [17]. Furthermore, doctors can cross-reference their records with existing databases to obtain comparisons helpful in determining trends in prescriptions and treatment.

Despite the importance of subcategorization in TCM, there has been no previous work directly addressing this issue (see Related Work [section 6] for a review of the most relevant studies). In this paper, we conduct a study of this problem and aim to accurately compute the similarity between patient records in TCM, a key step toward accurate subcategorization.

Computationally, the problem of subcategorization can be solved by applying a clustering algorithm to group similar records together and then induce subcategories. Indeed, such a clustering approach is also the basis of our study. However, a straightforward application of a clustering algorithm for subcategorization in which we directly match two patient records is unlikely to be effective because it does not address the issue of variations in both symptoms and herbs (i.e., comorbid symptoms and functionally similar herbs). For example, yellow tongue coating (黄苔) and greasy tongue coating (滑苔) are comorbid symptoms, and often appear together for patients with the “excess heat” (上火) syndrome. The crow-dipper (半夏) and the Chinese goldthread (黄连) are functionally similar herbs, and are commonly prescribed to treat vomiting and abdominal pain.

Moreover, TCM patients typically have ten to fifteen symptoms and are prescribed a similar number of herbs, further complicating subcategorization. We conclude that we must be able to match patient records “inexactly” by tolerating variations in both symptoms and herbs, yet allowing related symptoms or herbs to somehow match with each other. This is the main technical challenge that we address in this paper.

We propose to leverage the known herb-symptom associations in a TCM dictionary of herb functions [6] to discover latent relationships induced by comorbid symptoms and functionally similar herbs. These relationships enable inexact matchings of both symptoms and herbs, thereby improving the quality of subcategorization. Specifically, we can construct a network to represent the associations of symptoms and herbs in the dictionary as a graph.

Then, we apply a network embedding approach to learn vector representations for all of the symptoms and herbs while preserving their original associations in the dictionary. The vector representations allow us to compute similarities between any pair of symptoms or herbs. With these similarity scores, patient records that have no overlapping symptoms or herbs might partially match if the symptoms or herbs are related according to the knowledge in the dictionary. We call this subcategorization approach using dictionary-based embedding PaReCat (Patient Record Subcategorization).

Our hypothesis is that the prior knowledge of herb-symptom associations in the TCM dictionary can be used to discover latent relationships induced by comorbid symptoms and functionally similar herbs, thereby improving the quality of subcategorization. We performed extensive experiments on large-scale real-world datasets. As expected, this leads to more accurate matchings of patient records than baseline approaches, and thus better subcategorization results. We also show that PaReCat can be used immediately in multiple TCM clinical applications, such as retrieving similar patients as well as discovering meaningful cases of similar symptoms treated by different herbs and different symptoms treated by similar herbs.

2. PROBLEM DEFINITION

PaReCat takes two entities as input. The first is a set
of $n$ TCM patient records, $R = \{r_1, \ldots, r_n\}$, where $r_i \in R$ is a patient record and consists of $H = \{h_1, \ldots, h_p\}$ and $S = \{s_1, \ldots, s_l\}$. Here, $H$ is the set of herbs and $S$ is the set of symptoms for patient $i$. The second input is a set of known herb-symptom associations available in the TCM dictionary, which are functions of herbs outputting sets of symptoms, $f : H \rightarrow S$. An herb might have zero, one, or multiple symptom associations. PaReCat outputs a set of patient record categories $C = \{c_1, \ldots, c_m\}$, where each category $c_j \in C$ contains a subset of $R$.

We distinguish two kinds of categorizations. The first aims to group patients by considering both symptom and herbs. This allows us to create subcategories useful for cross-referencing patient treatments and discovering prescription trends. The second involves using only symptoms. This type of categorization emulates situations in which doctors are presented with a new patient and would like to view similar ones. We do not consider categorizations using only herbs because doctors do not prescribe treatments without first identifying symptoms.

Desirable categories will contain patients that are closely related by their symptoms, prescribed herbs, or both. These categories can be obtained by grouping similar patient records together. However, simply clustering TCM patients by their symptoms and prescriptions is ineffective. One cause for this difficulty is that TCM patients may display comorbid symptoms or be prescribed functionally similar herbs, but have very different ailments. These situations occur because many symptoms, such as coughing or bleeding, are not specific to any one disease.

In a similar vein, when prescribing treatments to patients, TCM doctors will assign batches of herbs, some of which are intended to treat the underlying syndromes instigating the observed symptoms. As a result, two patients who suffer from different diseases may be prescribed very similar herbs if they have similar syndromes. For example, a patient suffering from diarrhea was prescribed eleven herbs, seven of which were also prescribed to an asthma patient (Table 1).

Conversely, TCM patients might have identical diagnoses, yet very different herb prescriptions. This happens because the patients belong to different subcategorizations. For instance, between two herbs $A$ and $B$ that treat the same disease, a doctor might opt to prescribe $A$ over $B$ to a patient with chronic gastritis because the patient displays the “excess heat” (上火) syndrome, which $A$ specifically treats, but $B$ does not. In a particular case in our dataset, a patient suffering from asthma was prescribed 66 different herbs, none of which overlapped with another asthma patient’s prescription of 57 herbs (Table 2).

These circumstances occur only in TCM and do not arise in western medical cases, in which doctors prescribe only a handful of drugs that each treat specific symptoms or diseases. Systematic discovery of the same symptoms treated by variations of herbs and variations of symptoms treated by the same herbs is essential in transforming empirical TCM data into useful medical knowledge. This is the main goal of our study. PaReCat aims to capture these complications by leveraging a dictionary containing prior TCM knowledge.

### 3. METHODS

In this section, we first give an overview of the model’s...
pipeline (Figure 2), then discuss each component in detail.

3.1 A Brief Overview of PaReCat

We start by obtaining a set of known herb-symptom associations in the dictionary [6], which contains rules that map herbs to the symptoms they treat. For example, the “crow-dipper” has multiple entries, treating symptoms from vertigo to breathing difficulties. From these associations, we can construct a bipartite network in which one part of the network consists of symptoms and the other part consists of herbs. Symptoms that are associated with similar herbs will be close to each other in the network, and vice versa.

We then apply the network embedding approach to learn a low-dimensional vector representation for each herb and symptom. These low-dimensional vectors optimally preserve the original associations between symptoms and herbs. We can then compute the similarity between each pair of features by using the cosine similarity between their corresponding low-dimensional vectors. By using these similarity scores, which enable inexact matchings of symptoms and herbs, we can determine the similarity between any two patients, even if they share no herbs or symptoms.

Finally, we apply agglomerative clustering on the patient record and learn the cluster for each patient. The main novelty of our method lies in utilizing external knowledge (the herb-symptom dictionary) to cluster TCM patient records.

3.2 Building the Bipartite Network

First, from the set of known herb-symptom associations \( A : H \rightarrow S \), we construct a bipartite network \( G = (H, S, E) \), in which the two disjoint sets, \( H \) and \( S \), are the sets of herbs and symptoms, respectively. For each mapping \( h \rightarrow s (h \in H, s \in S) \) in \( A \), we add a node for \( h \) and a node for \( s \) if they are not already in \( G \), and create an edge between them. Thus, \( |E| = |A| \). \( G \) is not necessarily connected.

3.3 Network Embedding

Next, we perform network embedding, which takes as input the bipartite network \( G \). From \( G \), we use a recently developed network embedding approach, diffusion component analysis (DCA), to learn low-dimensional vector representations of the herbs and symptoms in the network [4]. DCA has been shown to achieve state-of-the-art results in learning network structure for gene function prediction [20].

DCA takes a network as input and outputs a low-dimensional representation for each node in the network. It ensures that two nodes will have very similar low-dimensional representations if they are topologically close in the network. Thus, related symptoms and related herbs tend to have similar low-dimensional vector representations, enabling inexact matchings. Between each pair of vectors, we compute the cosine similarity score to find the association between the corresponding symptoms or herbs. Each symptom and herb will have a similarity score with every other symptom and herb. For example, \( \text{Streptococcus dysgalactiae} \) (a bacteria strain that causes indigestion) and \( \text{yinianjin} \) (a powder consisting of several herbs) have a similarity score of 0.85272. Indeed, a capsule with \( \text{yinianjin} \) as the primary ingredient has been developed to treat children afflicted with \( S. \text{dysgalactiae} \) [2].

We compute a total of \( (|H| + |S|)^2 \) similarity scores. The similarity score between a feature and itself will always equal 1. With these scores, we output a similarity matrix \( M \) of shape \((|H| + |S|) \times (|H| + |S|)\). The diagonal consists of only 1’s, ensuring that an exact matching is always treated as the most reliable match. Non-zero values off the diagonal are in the range of \([-1, 1]\). These values capture the extent to which different symptoms and herbs match with each other. Furthermore, we set a similarity score threshold \( s \), and only consider entity pairs with a similarity score larger than \( s \). Scores in \( M \) below \( s \) are changed to 0.

For any particular run of PaReCat, the set of features can be symptoms or both symptoms and herbs. As a standard step in text mining, we found that filtering out high-frequency words and low-frequency words leads to better results. Thus, we also introduce parameters to remove herbs that appear in fewer than \( \beta_{\text{herbs}} \) patient visits (\( \beta_{\text{herbs}} \in \mathbb{Z} \)) and more than \( \alpha_{\text{herbs}} \) of all patient visits (\( 0.0 \leq \alpha_{\text{herbs}} \leq 1.0 \)). Our model similarly removes symptoms for \( \alpha_{\text{symptoms}} \) and \( \beta_{\text{symptoms}} \). If the features are only symptoms, then the values of \( \alpha_{\text{herbs}} \) and \( \beta_{\text{herbs}} \) are inconsequential.

In the end, we obtain an \( n \times n \) similarity matrix \( M' \) where \( n \leq |H| + |S| \) is the number of symptoms and herbs post-processing. Note that the inexact matchings allowed by matrix \( M' \) surpass symptom-symptom and herb-herb matchings to also allow symptom-herb matchings.

3.4 Enriching the Patient Matrix

After learning the similarity matrix \( M' \) from the dictionary, we use it to enrich the \( p \times n \) patient matrix \( P \), where \( p \) is the number of patients in the medical record. Each row of \( P \) is a binary vector corresponding to a patient, where a 1 denotes that the patient is prescribed the herb or possesses the symptom corresponding to the column, 0 otherwise. \( n \) is the number of symptoms and herbs, as before. The purpose of this enrichment is to make some zero elements (absent symptoms or herbs) in the patient matrix non-zero if they have enough support from known associated symptoms and herbs. Support for an element is based on the similarity matrix \( M' \). Formally, we perform the following matrix multiplication:

\[
P' = P \cdot M'
\]

where \( P' \) is the enriched \( p \times n \) patient record matrix in which each patient is regarded as a sample, and each herb or symptom a feature.

Intuitively, after the matrix multiplication, symptoms and herbs that have high similarity to many other symptoms and herbs in the original patient vector will tend to have higher values. In effect, this augments the original patient vector to potentially include additional related symptoms and herbs.
3.5 Agglomerative Clustering

Lastly, we cluster on the enriched patient vectors, which accommodate inexact matchings and enable more accurate patient record matching. We can use any similarity function to compute the similarity between two patient records. In our experiments, we chose cosine similarity as the affinity measure in the clustering, which was utilized in the similar task of western medical record linkage [15]. Similarly, our enriched vector representation can also support any clustering algorithm. In our experiments, we chose agglomerative clustering with average linkage, which has been shown to be useful for clustering herbs in TCM data [9]. The advantage of such a hierarchical clustering algorithm is that we can obtain a detailed subcategorization at different levels.

4. EXPERIMENTS

In this section, we discuss the evaluation of PaReCat. First, we describe the datasets used in this paper. Next, we provide the values of the tuned parameters. Lastly, we discuss both our quantitative and qualitative model evaluation.

4.1 Data Description and Experimental Setting

We conduct experiments on two datasets to quantitatively and qualitatively evaluate the effectiveness of PaReCat. After obtaining clusters of patient records, we evaluate with disease labels in the patient records as the ground truth: records that have the same label are expected to be in the same cluster.

4.1.1 Medical Textbook

We use a TCM textbook containing 2,276 patient medical records. These patient medical records are organized into a three-level hierarchy based on their identifying categories (Figure 3). There are three categories at the topmost level, which we refer to as level 1: “women and children,” “internal medicine,” and “surgical acupuncture.” The middle level, which we call level 2, is more specific, and includes categories such as “exogenous seasonal diseases.” The bottommost level, or level 3, includes even more specific ailments such as “the common cold.”

We refer to categories as labels for their corresponding patient records. There are three level 1 labels, 51 level 2 labels, and 274 level 3 labels. Because we can interpret these labels as the ground truth, this dataset is ideal for quantitative evaluation.

4.1.2 Stomach Disease Patient Records

In addition to the medical textbook, we further evaluate our model on a much larger medical record containing 9,529 anonymous patients, obtained from a major hospital in China. These patients all had some variety of stomach disease. This dataset does not have detailed labels for quantitative evaluation. We use it to understand whether our approach can offer interesting insights into stomach disease subcategories. The doctor who treated these patients manually assessed the subcategorization results.

4.1.3 Setting the Parameters

In addition to agglomerative clustering, we employed k-means, spectral clustering, and agglomerative clustering are shown in Table 3. For clustering on symptom features also used αsymptoms = 0.05 and βsymptoms = 1. For clustering without embedding on symptom and herb features, we obtained the best results with αsymptoms = 0.2, βsymptoms = 5, αherbs = 0.1, and βherbs = 2. To strengthen our baseline, we fixed these optimal parameters for their corresponding counterparts with embedding (i.e., clustering with embedding on symptom features also used αsymptoms = 0.05 and βsymptoms = 1). Tuning s for clustering with embedding, we achieved best results for symptom features with s = 0.96 and for combined symptom and herb features with s = 0.98.

4.1.4 Evaluation Metrics

We score the quality of the clusters with the adjusted Rand index [10]. It has been widely used in evaluating clustering results [18], defined as

\[ ARI = \frac{\sum_{i,j} (n_{ij}^2) - [\sum_i (n_i^2) \sum_j (n_j^2)]^{1/2}}{\frac{1}{2}[\sum_i (n_i^2)^2 + \sum_j (n_j^2)^2] - [\sum_i (n_i^2) \sum_j (n_j^2)]^{1/2}} \]

where \( n_{ij}, a_i, \) and \( b_j \) are values from the contingency table generated from the overlap between two groupings \( X = \{X_1, \ldots, X_r\} \) and \( Y = \{Y_1, \ldots, Y_s\} \). Each entry \( n_{ij} \) denotes the number of objects in common between \( X_i \) and \( Y_j : n_{ij} = |X_i \cap Y_j| \). \( a_i \) denotes the sum of the entries for the group \( X_i \), and \( b_j \) denotes the sum of the entries for the group \( Y_j \). A higher score indicates a better clustering with respect to the ground truth. The adjusted Rand index is in the range of \((-\infty, 1]\).
Table 4: In this subcategorization example, all patients have qi (气) or yin (阴) deficiency syndromes.

<table>
<thead>
<tr>
<th>disease</th>
<th>symptoms</th>
<th>herbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>chronic gastritis</td>
<td>weakness (乏力), pale tongue (舌淡), crenated tongue (舌边有齿痕), yellow tongue coating (苔薄黄), deep, thready pulse (脉沉细)</td>
<td>A. propinquus (黄芪), bugbane (开蕨), Chinese goldthread (黄连), coixseed (薏苡仁), crow-dipper (法半夏), female ginseng (当归), fried white atracytyldoles (麸炒白术), ginger (砂仁), gingederm magnolia-bark (姜厚朴), mandarine peel (陈皮), Mongolian dandelion (蒲公英), P. arecae (大腹皮), poor man’s ginseng (党参), sweet woodruff (青蒿), tuckahoe (茯苓)</td>
</tr>
<tr>
<td>chronic gastritis</td>
<td>weakness (乏力), weight loss (纳差), hiccup (干哕), dry mouth (口干), recurrent oral ulcers (口腔溃疡反复发作), joint pain (关节疼痛), lack of sleep (睡眠不实), pale tongue (舌淡), tongue lacerations (舌面有裂痕), yellow tongue coating (苔薄黄), deep, thready pulse (脉沉细)</td>
<td>A. asphodeloides (知母), A. propinquus honey (蜜制黄芪), Baikal skullcap (黄芩), calamus (菖蒲), Chinese figwort (玄参), Chinese liquorice honey (甘草), crow-dipper (法半夏), female ginseng (当归), fried A. lancea (麸炒苍术), fried barley (麦芽), fried Chinese peony (炒白芍), fried jujube (麸炒枣仁), fried white atracytyldoles (麸炒白术), gypsyffibrosinum (石斛), Mongolian dandelion (蒲公英), pine silk tree (合欢皮), poor man’s ginseng (党参), sweet woodruff (青蒿), tuckahoe (茯苓)</td>
</tr>
<tr>
<td>chronic gastritis</td>
<td>abdominal pain (上腹隐痛), acid reflux (反酸), halitosis (口中异味), dry mouth (口干), poor sleep (眠差), hair loss (脱发), dysmenorrhea (痛经), pale tongue (舌淡), crenated tongue (舌边有齿痕)</td>
<td>A. propinquus (黄芪), Baikal skullcap (黄芩), bitter orange (枳壳), black sesame (黑芝麻), Chinese knotweed (首乌藤), Chinese liquorice honey (甘草), Chinese thuja leaves (生侧柏叶), Chinese tree peony (牡丹皮), curcuma (郁金), female ginseng (当归), fried jujube (麸炒枣仁), fried white atracytyldoles (麸炒白术), R. glutinosus (生地黄), red sage (丹参), tuckahoe (茯苓)</td>
</tr>
<tr>
<td>hiatal hernia</td>
<td>abdominal distension (腹胀), acid reflux (反酸), belching (嗳气), dry mouth (口干), weakness (乏力), dry stool (大便干), poor sleep (睡眠欠佳)</td>
<td>A. propinquus (黄芪), Baikal skullcap (黄芩), bitter orange (枳壳), Chinese goldthread (黄连), Chinese liquorice honey (甘草), crow-dipper (法半夏), cuttlebone (海螵蛸), F. thunbergii (浙贝母), female ginseng (当归), fried Chinese peony (炒白芍), gingederm magnolia-bark (姜厚朴), immature bitter orange (枳实), Japanese Inula (旋覆花), perilla stem (紫苏梗), poor man’s ginseng (党参), red thorowax (柴胡), redstem wormwood (茵陈), T. ruticarpum (制吴茱萸), turmeric rhizome (姜黄)</td>
</tr>
</tbody>
</table>

clustering without embedding (first three rows), we note that agglomerative clustering performs the best, as expected.

The bottom two rows of the table show the results comparing agglomerative clustering without and with embedding (baseline and PaReCat, respectively). We observe improvement by embedding known herb-symptom associations for both types of features. Embedded feature vectors of symptoms and herbs achieved an adjusted Rand index of 0.2754, which is higher than that of feature vectors without embedding, 0.2717. Similarly, embedded feature vectors of symptoms alone achieved an adjusted Rand index of 0.1672, an improvement on that of feature vectors without embedding, 0.1613. The clustering results were very sensitive to s. For features of both symptoms and herbs, decreasing s to 0.9 lowered the adjusted Rand index of the resulting clusters to 0.2214.

Though the improvement is small, we have shown that PaReCat can indeed improve clustering performance by adding external information. Furthermore, we can gain new knowledge from clusters that mismatch the ground truth.

4.1.6 Qualitative Evaluation

To investigate whether PaReCat can effectively cluster patients into informative subcategories, we apply PaReCat to the larger set of stomach disease medical records. Here, we randomly selected three clusters and had the TCM doctor who treated the patients verify PaReCat’s effectiveness. The expert found two of them to be especially coherent and informative (Tables 4 and 5, where each row denotes a patient record).

In the first cluster (Table 4), all of the patients were diagnosed with chronic gastritis. In addition to this similarity, the records state that these patients showed syndromes of either qi deficiency (气虚) or yin deficiency (阴虚). The related symptoms here are weakness (乏力) and pale tongue (舌淡) for qi deficiency (气虚), dry mouth (口干) and dry stool (大便干) for yin deficiency (阴虚), and deep, thready pulse (脉沉细) for both deficiencies. The corresponding herbs for these patients specifically target qi deficiency (气虚) or yin deficiency (阴虚). For example, poor man’s ginseng (党参), A. propinquus honey (蜜制黄芪), tuckahoe (茯苓), and fried white atracytyldoles (麸炒白术) strengthen the spleen (益气健脾), R. glutinosus (生地), female ginseng (当归), and Chinese peony (白芍) enrich blood (滋阴养血). These benefits supplement and fortify the qi (气) and yin (阴). The symptoms and herbs of interest are bolded in the table. This example illustrates how TCM treats its patients by identifying the underlying syndromes.

We show another meaningful cluster in Table 5. The records state that all of the patients had symptoms associated with the “excess heat” (上火) syndrome, such as coughing (咳嗽), chest tightness (胸闷), dry mouth (口干), and dry stool (大便干). Consequently, the patients were prescribed herbs specifically for these symptom-syndrome combinations. We can see that there are not many overlaps in symptoms and herbs among the patients, which may be the
Table 5: In this subcategorization example, all patients have symptoms associated with the “excess heat” (上火) syndrome.

<table>
<thead>
<tr>
<th>disease</th>
<th>symptoms</th>
<th>herbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>chronic gastritis</td>
<td>acid reflux, heartburn (烧心), dry mouth (11 F), halitosis (口臭), belching (嗳气), excess phlegm (痰多), poor appetite (食欲欠佳), coughing (咳嗽), dry stool (大便干), cracked tongue (舌面有裂纹)</td>
<td>red sage (丹参), crow-dipper (法半夏), monkeygrass (麦冬), umbrella polypore (猪苓), magnolia-bark (厚朴花), A. propinquus honey (蜜黄芪), Oriental water-plantain (泽泻), perilla leaf (紫苏叶), Chinese liquorice honey (蜜甘草), yam extract (山药), Chinese cornel dogwood (酒山萸肉), American silvertop (北沙参), R. glutinosa (地黄), gypsum fibrosum (石膏), Lophatherum (淡竹叶), Chinese gourd (瓜篓), bishop’s weed (鱼腥草), cinnamon (肉桂), tuckahoe (茯苓)</td>
</tr>
<tr>
<td>chronic gastritis</td>
<td>acid reflux (反酸), coughing (咳嗽), white sputum (咳痰色白), dry stool (大便干), vomiting (呕吐), hypouresis (小便量少)</td>
<td>A. chinensis (藤藜根), nacre (大贝母), cuttlebone (乌贼骨), tuckahoe (茯苓), Chinese goldthread (黄连), peach seeds (桃仁), Chinese liquorice (甘草), Baikal skullcap (黄芩), crow-dipper (法半夏), Chinese peony (白芍), gingered magnolia-bark (姜厚朴), Chinese rhubarb (生大黄), citron fruit (香橼), bitter orange (枳壳), perilla stem (紫苏梗)</td>
</tr>
<tr>
<td>chronic gastritis</td>
<td>abdominal pain (上腹疼痛), belching (嗳气), heartburn (烧心), acid reflux (反酸), constipation (大便干燥), coughing (咳嗽), dark, greasy tongue coating (舌苔白腻), thready pulse (脉沉细弦)</td>
<td>crow-dipper (法半夏), Chinese liquorice honey (蜜甘草), bitter orange (枳壳), Chinese goldthread (黄连), gingered magnolia-bark (姜厚朴), red thorowax (柴胡), bamboo extract (竹茹), Mongolian dandelion (蒲公英), fried Chinese peony (炒白芍), O. diffusa (白花蛇舌草), turneric rhizome (黄芩), cuttlebone (海螵蛸), Baikal skullcap (黄芩)</td>
</tr>
<tr>
<td>chronic gastritis</td>
<td>abdominal pain (上腹疼痛), fullness (痞满), coughing (咳嗽), dry mouth (11 F), dry stool (大便干), dark, swollen tongue (舌苔厚腻), thready pulse (脉沉细弦)</td>
<td>perilla leaf (紫苏叶), fried white atractyloides (炙炒白术), bugbane (升麻), Chinese bellflower (桔梗), red thorowax (柴胡), Baikal skullcap (酒黄芩), tuckahoe (茯苓), immature bitter orange (枳实), Chinese goldthread (黄连), ginger (干姜), Japanese Inula (旋复花), Java grass (醋香附), mandarine peel (陈皮), crow-dipper (法半夏), curcuma (郁金), gingered magnolia-bark (姜厚朴), cinnamon (肉桂), chicken gizzard (鸡内金), American silvertop (北沙参), bitter orange (枳壳), bishop’s weed (鱼腥草)</td>
</tr>
</tbody>
</table>

5. APPLICATIONS OF PARECAT

As mentioned in the previous section, PaReCat can be used to achieve high quality, informative subcategorizations of patients. Here, we show how we can use them in three different applications to directly support doctors.

5.1 Similar Patient Retrieval

PaReCat can be used to retrieve similar patients. We show the results of a sample retrieval here. We extract the five most similar pairs of patients as computed by PaReCat, of which we highlight two.

According to the doctor, the first pair of patients (Table 6) share the underlying syndrome of cold deficiency (虚寒), represented by symptoms such as body chills (全身畏寒) and cold hands and feet (手足冷). In addition, they also showed symptoms for the heat (热) syndrome, such as dry mouth (口干), bitter taste (口苦), halitosis (口中有物), and yellow tongue coating (舌苔黄). These two patients’ conditions are typical examples of cold and heat (中医寒热错杂) syndromes. As a consequence, doctors prescribed both cold (寒) and heat (热) treatment herbs.

The second pair of patients are both diagnosed with upper respiratory tract infection (URTl, 上呼吸道感染) in addition to chronic gastritis (慢性胃炎) (Table 7). Recall that our method does not utilize the disease diagnosis information in clustering. However, we still manage to find these patients, despite the fact that they share no symptoms. Among the herbs shared by the two patients, four of them (powdered water buffalo horn (水牛角粉) [23], Japanese apricot (乌梅) [8], mint (薄荷) [12], and wood root (板蓝根) [13]) are specific treatments for URTl. Our methods successfully identified these four herbs, clustering the two patients together.

5.2 Similar Symptoms with Different Herbs

Another application is identifying patients that have similar symptoms but are treated with different herbs. Conversely, identifying patients with different symptoms but treated with similar herbs is also an interesting task. These cases are unique to TCM, since western medicine tends to treat the same symptoms with the same drugs. Our method is able to identify these two categories of patients.

The first case (similar symptoms, different herbs) is especially important, and plays a large role in TCM misdiagnoses. An inexperienced doctor may see a patient with symptoms similar to one he or she had previously treated. Traditional methods might confound patients with similar symptoms. However, PaReCat can view previous medical records, taking into account both symptoms and herbs, and help new doctors improve decision-making when facing similar circumstances.

For example, a pair of patients in our medical record both displayed dry mouth (口干), bloating (胃胀), insomnia (睡眠多梦), acid reflux (反酸), and a yellow tongue coating (舌苔黄) (Table 8). However, one patient is treated with Oldenlandia diffusa (白花蛇舌草) and vinegarized chicken gizzard (醋鸡内金). This patient suffered from gallbladder polyps, which the herbs specifically treat (O. diffusa [11], vinegarized chicken gizzard [7]). On the other hand, the other patient was not treated with these herbs, and was instead diagnosed with superficial gastritis (浅表性胃炎).
### Table 6: Example of a pair of similar patients with cold and heat (中医寒热错杂) syndromes.

<table>
<thead>
<tr>
<th>disease</th>
<th>symptoms</th>
<th>herbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>chronic gastritis (慢性胃炎)</td>
<td>body chills (全身畏寒), dry mouth (干口), heat excess (易上火), cold hands and feet (手足凉), yellow complexion (面黄), watery stool (大便不干), bitter taste (苦), glossodynia (舌痛), halitosis (口臭)(喜酸), dark red, swollen tongue (舌暗红胖), yellow tongue coating (舌苔黄腻)(根厚)</td>
<td>poor man's ginseng (党参), fried white atractylodites (炒炒白术), Chinese cinnamon (桂枝), Chinese liquorice (甘草), Chinese goldthread (黄连), <em>R. glutinosa</em> (生地黄), shrubby sophora (苦参), gypsum fibrosum (石膏), common rush (灯心草), fried A. lancea (麸炒苍术)</td>
</tr>
<tr>
<td>chronic gastritis (慢性胃炎)</td>
<td>stomach pain (胃痛), cold hands and feet (手足凉), body chills (全身畏寒), dry mouth (干口), halitosis (口臭), dizziness (头晕), watery stool (大便不干), bitter taste (苦), sore gums (牙龈疼痛), dark red, swollen tongue (舌暗红胖), yellow tongue base coating (舌苔黄腻根厚)</td>
<td><em>A. propinquus</em> (黄芪), fried white atractylodites (麸炒白术), kudzu (葛根), Chinese goldthread (黄连), Chinese liquorice honey (蜜甘草), fried gardenia (炒栀子), red thorowax (柴胡), Chinese tree peony (牡丹皮), bamboo extract (竹茹), <em>R. glutinosa</em> (生地黄), Mongolian dandelion (蒲公英), fried Chinese peony (炒白芍), gypsum fibrosum (石膏), Baikal skullcap (黄芩), <em>A. asphodeloides</em> (知母), <em>R. reticulatum</em> (制吴茱萸), female ginseng (当归), Chinese rhubarb (生大黄)</td>
</tr>
</tbody>
</table>

### Table 7: Example of a pair of similar patients with no common symptoms.

<table>
<thead>
<tr>
<th>disease</th>
<th>symptoms</th>
<th>herbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>chronic gastritis (慢性胃炎), URTI (上呼吸道感染)</td>
<td>cracked tongue (舌苔裂纹), pale tongue (舌淡), bloating (胃胀), stomach pain (胃痛), weakness (乏力), epigastric chills (胃脘提寒), white tongue coating (苔薄白), belching (暖气), reddish tongue (舌偏红), epigastric pain (胃脘隐痛), bowel discomfort (肠鸣), abdomen chills (腹部畏寒)</td>
<td>ginger (砂仁), Chinese liquorice honey (蜜甘草), mint (薄荷), crow-dipper (法半夏), Chinese tree peony (牡丹皮), Chinese cinnamon (桂枝), <em>powdered water buffalo horn</em> (水牛角粉), fried white atractylodites (麸炒白术), shrubby sophora (苦参), fried coixseed (炒薏苡仁), weeping forsythia (连翘), bitter orange (枳壳), tuckahoe (茯苓), <em>Japanese apricot</em> (乌梅), Chinese goldthread (黄连), American silvertop (北沙参), hyacinth orchid (白及), <em>wood root</em> (板蓝根), Chinese parsnip root (防风), lesser reedmace (蒲黄), Baikal skullcap (黄芩), fried Chinese peony (炒白芍), Japanese honeysuckle (金银花), poor man’s ginseng (党参), costus (木香), female ginseng (当归), <em>Japanese apricot</em> (乌梅), Chinese goldthread (黄连), Baikal skullcap (黄芩), fried Chinese peony (炒白芍), female ginseng (当归), costus (木香), ginger (砂仁), Chinese tree peony (牡丹皮), <em>powdered water buffalo horn</em> (水牛角粉), shrubby sophora (苦参), American silvertop (北沙参), <em>wood root</em> (板蓝根), Chinese parsnip root (防风)</td>
</tr>
<tr>
<td>insomnia (失眠)</td>
<td>dry mouth (口干), thready pulse (脉沉弦), thirst (喜饮), dry throat (咽干), dry lips (唇干), greasy tongue coating (苔白腻), swollen, crenated tongue (舌苔边有齿痕)</td>
<td>Chinese liquorice honey (蜜甘草), crow-dipper (法半夏), mint (薄荷), weeping forsythia (连翘), bitter orange (枳壳), tuckahoe (茯苓), <em>Japanese apricot</em> (乌梅), Chinese goldthread (黄连), Baikal skullcap (黄芩), fried Chinese peony (炒白芍), female ginseng (当归), costus (木香), ginger (砂仁), Chinese tree peony (牡丹皮), <em>powdered water buffalo horn</em> (水牛角粉), shrubby sophora (苦参), American silvertop (北沙参), <em>wood root</em> (板蓝根), Chinese parsnip root (防风)</td>
</tr>
</tbody>
</table>

### Table 8: Two patients clustered together, with similar symptoms, but have different herbs treating different diseases.

<table>
<thead>
<tr>
<th>disease</th>
<th>symptoms</th>
<th>herbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>superficial gastritis (浅表性胃炎)</td>
<td>dry mouth (干口), dry stool (大便干), heartburn (烧心), halitosis (口臭), insomnia (睡眠时有欠佳), abdomen chills (腹部畏寒), belching (嗳气), bloating (胃胀), yellow tongue coating (苔黄), acid reflux (反酸), bitter taste (口苦)</td>
<td>Chinese goldthread (黄连), Chinese liquorice honey (蜜甘草), sandalwood (檀香), ginger (砂仁), crow-dipper (法半夏), cuttlebone (海螵蛸), immature bittersweet orange (枳实), female ginseng (当归), ginger (干姜), Chinese gourd (瓜蒌), red sage (丹参), <em>Java grass</em> (醋酸香), bitter orange (枳壳), Baikal skullcap (黄芩), <em>M. toosendan</em> (川楝子)</td>
</tr>
<tr>
<td>gallbladder polyps (胆囊息肉)</td>
<td>dry mouth (干口), thready pulse (脉弦数), insomnia (睡眠多梦), bloating (胃胀), acid reflux (反酸), yellow tongue coating (苔黄腻), frequent urination (尿频), cracked tongue (舌裂纹)</td>
<td><em>O. diffusa</em> (白花蛇舌草), immature bittersweet orange (枳实), Chinese gourd (瓜蒌), cuttlebone (海螵蛸), bitter orange (枳壳), Baikal skullcap (黄芩), monkeygrass (麦冬), Chinese figwort (玄参), ginger (干姜), fried Chinese peony (炒白芍), female ginseng (当归), mandarin peel (陈皮), crow-dipper (法半夏), gingered magnolin-bark (姜厚朴), Chinese cinnamon (桂枝), Chinese rhubarb (生大黄), <em>vinegared chicken gizzard</em> (醋鸡内金)</td>
</tr>
</tbody>
</table>
5.3 Different Symptoms with Similar Herbs

In addition to patients with similar symptoms but different herbs, it is interesting to study patients that have different symptoms but are treated with similar herbs. Table 9 shows two patients that have very different symptoms. However, they are prescribed similar herbs. M. toosendan (川楝子) is a particular herb of note, which specifically treats liver issues [22]. Indeed, both patients are diagnosed with liver ailments (hepatic steatosis and cirrhosis). This relationship is not uncommon to TCM; doctors often prescribe a multitude of herbs as a supplement to a main herb (in this case, M. toosendan (川楝子)). PaReCat can successfully filter out ambiguous herbs to discover these relationships.

6. RELATED WORK

To the best of our knowledge, our method is the first to combine patient record analysis and herb-symptom associations for subcategorizing traditional Chinese medicine records. However, data-driven approaches on TCM data have attracted more attention in recent years. In one particular study, the authors employed Chi-Squared Automatic Interaction Detection (CHAID) decision trees to identify and differentiate syndromes associated with coronary heart disease, performing classifications on the syndromes with k-core network analysis [16]. One study combined network construction and cluster analysis to study relationships and associations among TCM patient records [19]. However, it performed analysis with only symptoms and did not leverage herb-symptom associations to solve the issues associated with comorbid symptoms and functionally similar herbs in TCM. He et al. also used agglomerative clustering on TCM data, but categorized herbs based on their efficacies in order to analyze their chemical components [9]. Roque et al. used cosine similarity between patient records to analyze disease co-occurrence [15]. However, they did not consider treatment information, which is crucial to patient record subcategorization. Furthermore, methods designed for western medical records are difficult to apply to TCM records because firstly, some symptoms (e.g., vacuity/depletion (虛/ 空)) are unique to TCM [21]. Secondly, TCM generally has more symptoms and herbs per patient. Our experimental results confirmed that the methods proposed to cluster western medical records indeed perform poorly on TCM datasets.

7. CONCLUSION

Mining subcategorizations from TCM medical records is an important task for precision medicine. In this paper, we proposed a novel patient record subcategorization model called PaReCat. PaReCat is able to obtain patient subcategorizations that characterize the underlying syndromes behind the observed symptoms and herbs. It uses a novel dictionary-based embedding approach to solve challenges associated with comorbid symptoms and functionally similar herbs. We performed experiments on two real-world datasets and observed substantial improvement in patient subcategorizations created by PaReCat. We have also verified the subcategorizations to be meaningful for understanding variations of stomach diseases. PaReCat’s generality allows it to be applied to any TCM dataset to discover interesting subcategories that are immediately useful to not only research, but also clinical applications. In the future, we plan to pursue this direction by applying PaReCat to the massive datasets that we have access to via the TCM Data Center in Beijing. PaReCat is completely unsupervised, which has the advantage of requiring no manual work. However, to improve the accuracy of subcategorization, we can explore semi-supervised subcategorizations in which we allow a doctor to provide feedback on the clustering results, which can then be used as supervision data for additional clustering.

To further improve our method, we can employ natural language analysis. One issue in the data that could be solved by natural language analysis is that textual differences separated essentially identical herbs and symptoms. For example, a doctor might prescribe a crushed variant of an herb to a patient, creating a new herb in the record, though the main ingredient remains the same.

Yet another interesting future direction opened by our work is to further assess the effectiveness of a set of prescribed herbs for each specific subcategory of patients by performing natural language analysis on the patient records. Once we can measure effectiveness, we will be able to lever-

<table>
<thead>
<tr>
<th>disease</th>
<th>symptoms</th>
<th>herbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>hepatic steatosis (脂肪肝)</td>
<td>weakness (乏力), laryngitis (声音嘶哑), white tongue coating (苔薄白), dark, purple tongue (舌紫暗), dark, red tongue (舌暗红), deep, thready pulse (脉沉细)</td>
<td>ginger (砂仁), curcumins (郁金), vinegared C. yanhusuo (醋延胡索), cuttlebone (海螵蛸), red thorowax (醋柴胡), bitter orange (枳壳), E. ulmoides (生决明), monkeygrass (麦冬), R. glutinosa (生地黄), Baikal skullcap (黄芩), Chinese peony root (赤芍), M. toosendan (川楝子), lotus leaf (荷叶), vinegared Java grass (醋香附), female ginseng (当归), gardenia (栀子)</td>
</tr>
<tr>
<td>gall-bladder inflammation (胆囊炎)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cirrhosis (肝硬化)</td>
<td>weakness (乏力), pale tongue (舌淡胖), dizziness (头晕), back pain (背痛), abdomen chills (腹部寒塞), chest tightness (胸闷), weight loss (纳差), belching (伴暖气), blurred vision (视物模糊), chest pain (胸骨后疼痛)</td>
<td>Chinese liquorice honey (白术草), magnolia-bark (厚朴草), perilla leaf (紫苏叶), cuttlebone (海螵蛸), bitter orange (枳壳), Baikal skullcap (黄芩), Chinese goldthread (黄连), fried Chinese peony (炒白芍), ginger (干姜), vinegared Java grass (醋香附), fried A. lancea (麸炒苍术), female ginseng (当归), mandarine peel (陈皮), crow-dipper (法半夏), pachouli (藿香), vinegared C. yanhusuo (醋延胡索), T. ruticarpurn (制吴茱萸), false starwort (太子参), perilla stem (紫苏梗), M. toosendan (川楝子)</td>
</tr>
</tbody>
</table>

Table 9: Two patients clustered together with different symptoms, but treated with similar herbs. Shared herbs are bolded.
age the many interesting subcategories generated by PaRe-Cat as promising new hypotheses about variations of diseases. These hypotheses can help discover new knowledge about effective personalized treatment for different subcategories.

8. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1144245. This research was also supported by grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). Our work also received support from National Key Technology R&D Program grants 2013BA02B01 and 2013BAH06F03, National S&T Major Project of China grant 2013BAI13B04, and Special Research Project of TCMS by State Administration of Traditional Chinese Medicine grant 201407001.

9. REFERENCES