Abstract—Traditional Chinese medicine (TCM) can provide important complementary medical care to modern medicine, and is widely practiced in China and many other countries. Unfortunately, due to its empirical nature and history of trial and error, effective diagnosis and prescription methods are not well-defined. This setback results in a significant challenge in retaining, sharing, and inheriting knowledge among physicians. In this paper, we propose a new asymmetric probabilistic model for the joint analysis of symptoms, diseases, and herbs in patient records to discover and extract latent TCM knowledge. We base our model on the comprehensive evaluation of modern medicine and TCM-specific symptoms in addition to herb prescriptions for particular diseases. Experimental results on a large dataset demonstrate the effectiveness of the proposed model for discovering useful knowledge and its potential clinical applications.

I. 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 [1]. Consequently, patient diagnosis and treatment in real-world settings are extremely difficult. In traditional Chinese medicine (TCM), a style of alternative medicine widely used in China for thousands of years, personalized treatment is the core principle of clinical practice. TCM doctors prescribe mixtures of herbs tailored to each patient. On the other hand, western medicine utilizes many “one size fits all” over-the-counter drugs for common diseases.

In TCM clinical settings, doctors record symptoms that they observe in their patients. The symptoms are captured and described in information entities such as “night sweats” and “night fevers.” Some symptoms are unique to TCM, with no English analogue. For example, “shuo ma,” which is caused by an excess of one’s “heat element,” results in a rapid pulse. These symptoms require TCM theory to identify, and necessitate personalized treatment. The doctors then prescribe herbs based on the patient’s disease profile. This personalized medical process is then detailed in the patient’s medical record.

Unlike modern medicine, TCM is flexible with prescriptions, assigning multiple ingredients according to the patient’s particular symptoms. This flexibility applies to cases of even the most common diseases. The manner in which TCM observes an individual’s symptom patterns is reminiscent of precision medicine techniques. Because of this, TCM can complement modern precision medicine, which currently relies on molecular profiles [2]. Learning how doctors analyze symptoms and arrive at prescriptions can greatly increase our knowledge of medicine.

Unfortunately, due to the empirical nature of TCM, key knowledge about disease diagnoses based on TCM-specific symptoms and herb prescriptions is not well-documented. This creates a significant challenge to retain, share, and inherit knowledge of TCM [3]. This is exacerbated by the fact that patients often have comorbid medical conditions and symptoms, which introduces a host of complex, intricately connected factors [4]. In general, medical data is considered to be one of the most difficult domains for data mining. It becomes even more intractable in TCM because of the aforementioned herb prescription combinations.

Fortunately, some of these challenges can be addressed by analyzing large-scale TCM clinical patient data. With this vision, the TCM Data Center in Beijing has been collecting TCM electronic medical records (EMRs) and integrating them into a clinical data warehouse since 2007 [5]. The data warehouse has curated over 300,000 clinical cases and will acquire patient records from six hospitals in the following years. Each hospital has received roughly 3,000,000 patient visits annually for a decade. This data provides an unprecedented opportunity to apply powerful mining algorithms to discover useful knowledge. Indeed, several interesting studies have already succeeded in detecting herb combinations for disease treatment [6], latent symptom phenotype regularities [7], and herb-symptom relationships [8]. Li et al. showed that discovering knowledge from TCM clinical data is feasible and meaningful [9].

In pursuit of this vision, we analyze large-scale datasets
obtained from the TCM Data Center. We propose a novel probabilistic model to jointly analyze symptoms, diseases, and herbs and to discover patterns that reveal knowledge about the following:

1) Typical symptom groups associated with disease conditions. This knowledge can help doctors diagnose diseases.
2) Typical herb groups associated with disease conditions. This knowledge can help doctors prescribe herbs.
3) Correlations among symptom groups and herb groups. This knowledge can facilitate personalized prescriptions and reveal new insights into herb effectiveness.

Generative probabilistic models can model an entire dataset as a whole. This optimizes pattern discovery by using all of the available data. Such models have already been successfully used for text mining [10] and TCM data analysis [11], [12]. However, previous works did not model the asymmetric causal relations among symptoms, diseases, and herbs; the relations were instead inaccurately considered to be symmetric. In this paper, we propose a new probabilistic model that can represent the causal relations asymmetrically to enable more accurate discovery of medical knowledge. In particular, we are interested in disease-specific knowledge.

We first introduce a new pattern, which we call a disease profile. It consists of three elements: a disease, a multinomial distribution over symptoms for the disease, and a multinomial distribution over herbs for the disease. For example, the disease profile for chronic gastritis shows that its most common symptoms are “bloating” and “epigastric pain and chills,” and its most frequently prescribed herbs are “stir-fried immature bitter orange in bran” and “crow-dipper.”

To discover disease profiles from our dataset, we propose a generative probabilistic model with the profiles as model parameters. We can obtain maximum likelihood estimates for these profile parameters, which can be interpreted as knowledge discovered from the data. These profiles are not only meaningful by themselves, but are also useful as references for recommending prescriptions to TCM practitioners.

Experimental results on a large TCM EMR collection demonstrate the proposed model’s effectiveness. Our approach successfully discovered useful TCM disease profiles, which a TCM physician verified. In addition, our approach also shows promising results in symptom-herb relationship mining as well as in disease and herb prediction.

II. Problem Definition and Formulation

Our dataset contains three types of data: patient symptoms, physician disease diagnoses, and prescribed herbs. In this work, we only consider the modern disease diagnosis. In previous works, the correlations of these three fields have been analyzed homogeneously [11], [13] without treating diseases differently from symptoms or herbs. A key observation that motivated our work is that the causal relation among these three fields is asymmetric (Fig. 3).

Specifically, the disease information provides important connections between herbs and symptoms, effectively reducing the space complexity of all possible matchings between symptoms and herbs. Moreover, explicit incorporation of diseases into the mining process also enables direct discovery of symptom-disease and herb-disease associations. Both types of associations have important clinical applications, with the former aiding diagnosis and the latter aiding herb prescription.

To mine these associations, we define our computational problem as follows: the input consists of a set of 3-tuples \((S,D,H)\), where \(S\) is a subset of all possible symptoms, \(D\) is a subset of all possible diseases, and \(H\) is a subset of all possible herbs. The output consists of a set of \(r\) disease profiles \(\{p_1,p_2,\ldots,p_r\}\) (Fig. 1). Each profile corresponds to a disease \(d\in D\), and is characterized by two distributions. One distribution is over symptoms, \(Pr(s|d), s\in S\), reflecting the typical symptoms of disease \(d\). The other is over herbs, \(Pr(h|d), h\in H\), capturing the typical herbs prescribed for patients with disease \(d\).

III. Methods

In a patient record, a single patient possesses a set of symptoms, to which a physician assigns a set of diseases. However, within the patient record, it is unknown whether a given symptom \(s\in S\) is related to some assigned disease \(d_i\in D\) or another assigned disease \(d_j\in D\) (Fig. 2). Furthermore, a patient record contains a set of prescribed herbs. The disease that each herb is meant to treat is also
unknown. This challenge of symptom and herb separation occurs very frequently in our TCM patient records, in which each record has an average of 15 symptoms and 20 herbs. As a result, separating symptoms and herbs is necessary to optimally deduce their corresponding diseases and thus prescribe the appropriate treatments to the patient. In this paper, we propose to solve the challenge of symptom and herb separation by using an unsupervised generative probabilistic model.

At a high level, we wish to model the causal relations among symptoms, diseases, and herbs, illustrated in the left portion of Fig. 3. A patient’s disease triggers a particular set of symptoms. This prompts the doctor to prescribe the appropriate set of herbs. The proposed probabilistic model then mimics these causal relations: $Pr(s|d)$ models the symptoms of a disease $d$, while $Pr(h|d)$ models the prescribed herbs. Using a distribution to represent the causal relations captures uncertainty, which is inevitable given the lack of documented, precise medical knowledge in TCM. There are at least two simplifications of these models, dependent upon the assumptions that we make.

1) **Multinomial model.** In this model, we assume that the symptoms compete with each other. If one symptom has an increased probability of appearing, then it decreases the appearance probability of other symptoms. This assumption allows us to infer which symptom is most likely to appear, giving us two constraints: $\sum_{s \in S} Pr(s|d) = 1$ and $\sum_{h \in H} Pr(h|d) = 1$. From the perspective of a generative model, we assume that the observed symptoms are generated by sampling symptoms using $Pr(s|d)$. Similarly, the observed herbs are sampled using $Pr(h|d)$.

2) **Multi-Bernoulli model.** We assume that symptoms independently appear. Under this assumption, we have two constraints: $Pr(s = 0|d) + Pr(s = 1|d) = 1$ and $Pr(h = 0|d) + Pr(h = 1|d) = 1$. Each symptom is generated independently from its symptom distribution. In other words, we flip a biased coin with probability $Pr(s|d)$ to determine whether symptom $s$ should show up for disease $d$. Herbs are generated in the same way, but with $Pr(h|d)$.

Neither of these models entirely captures the actual complexity of causal relations, but they create an excellent reference point. Additionally, more sophisticated models inevitably introduce additional parameters, which would then require even more data points to make accurate estimations.

In this paper, we focus on the multinomial model in the hopes of discovering the most prominent symptoms and herbs for each disease, leaving the exploration of the multi-Bernoulli model as future work. We now give a more detailed description of the multinomial model.

Intuitively, we model how symptoms and herbs of a patient record are generated from their corresponding set of diseases by using the conditional distribution $Pr(\{s_1, \ldots, s_n\}, \{h_1, \ldots, h_m\}|\{d_1, \ldots, d_k\})$. In order to characterize this generative model, we introduce another distribution over diseases for patient record $t$, $Pr(d|t)$. $Pr(d|t)$ decides the primary disease responsible for generating a specific set of symptoms and prescriptions.

To generate a set of symptoms $S = \{s_1, \ldots, s_n\}$, we generate each symptom $s_i \in S$ independently. We first sample a disease with probability $Pr(d|t)$, and then sample from its corresponding symptom model $Pr(s|d)$. We repeat this process $n$ times to generate the $n$ symptoms. The herbs $\{h_1, \ldots, h_m\}$ are generated in a similar way, using $Pr(h|d)$ instead of $Pr(s|d)$.

The probability of patient $t$ having symptom $s$ and being prescribed herb $h$ with disease diagnoses $D_t = \{d_1, \ldots, d_k\}$ will then be:

$$Pr(s, h|t) = \sum_{d \in D_t} Pr(d|t) Pr(s|d) Pr(h|d) \quad (1)$$

If we view symptoms and herbs as “words,” our proposed model is similar to basic topic models such as PLSA [10], which has been successfully applied to the task of discovering word clusters from noisy text collections. However, an important difference is that in PLSA, each word $w \in W$ in document $d \in D$ belongs to an underlying cluster (topic). In contrast, we assume that all of the clusters (diseases) are observed. This assumption not only applies to real clinical situations, but also restricts the model and allows us to directly learn disease-specific knowledge. In effect, we are imposing an infinitely strong prior on latent topics such that they are strictly tied to certain diseases.

A previous model for a similar dataset assumed latent associations between symptoms and herbs [13]. As a result of this limitation, the learned associations are not ensured to be aligned with particular diseases, despite using diagnosis information. We hypothesize that our model can discover more accurate disease-specific knowledge than this older, related model. The results of our experiments confirm this hypothesis.

Because common symptoms and herbs frequently co-occur with many different diseases, they are not discriminative enough to accurately infer a specific disease. To alleviate this problem, we further define a background cluster $Pr(\theta|t)$. The background cluster generates these noisy symptoms and herbs, allowing the final disease cluster to be more informative. A
patient record in the collection can now be generated by the following mixture model:

\[
\Pr(s, h|t) = \left( \sum_{d \in D_t} \Pr(d|t) \Pr(s|d) \Pr(h|d) \right) + \Pr(\theta_B|t) \Pr(s|\theta_B) \Pr(h|\theta_B)
\]

where \(\Pr(s|\theta_B)\) and \(\Pr(h|\theta_B)\) are the background multinomial distributions over symptoms and herbs, respectively.

By fitting this mixture model to patient records, we can find the multinomial distributions that best explain our data. These distributions are associated with the symptoms and herbs that correspond to each disease. Under the definition of \(\Pr(s, h, t)\), the log-likelihood function of the patient record collection \(M\) is:

\[
\log \Pr(S, H|T) = \sum_{t \in T} \sum_{s \in S} \sum_{h \in H} I(t, s, h) \times \log \Pr(s, h|t)
\]

where \(I(t, s, h) = 1\) if patient \(t\) has symptom \(s\) and was prescribed herb \(h\), 0 otherwise. Our model has the following parameters: \(\Pr(d|t)\), \(\Pr(s|d)\), \(\Pr(h|d)\), \(\Pr(\theta_B)\), \(\Pr(\theta_B|t)\), \(\Pr(s|\theta_B)\), and \(\Pr(h|\theta_B)\). We use an expectation-maximization (EM) algorithm to perform maximum likelihood estimation of this latent variable model. The EM algorithm alternately performs an expectation (E) step and a maximization (M) step, and is guaranteed to converge [14]. It attempts to find the optimal parameters to explain the model. Because the parameter estimation for the background cluster and herb distributions is similar to that of a symptom distribution, we only give the update formula for a symptom distribution. In the E step, we compute the posterior probabilities for \(\Pr(d|s, h, t)\).

\[
\Pr(d|s, h, t) \propto P^n(d)P^n(t|d)P^n(s|d)P^n(h|d)
\]

In the M step, we re-estimate the parameters based on the posterior probabilities obtained in the E step.

\[
P^{n+1}(d) \propto \sum_{t, h, s} I(t, s, h)P^n(d|s, h, t)
\]

\[
P^{n+1}(t|d) \propto \sum_{h, s} I(t, s, h)P^n(s|d, h, t)
\]

\[
P^{n+1}(s|d) \propto \sum_{t, h} I(t, s, h)P^n(d|s, h, t)
\]

\[
P^{n+1}(h|d) \propto \sum_{t, s} I(t, s, h)P^n(d|s, h, t)
\]

IV. TCM DECISION SUPPORT APPLICATION

Note that our model is completely unsupervised, so it requires neither human annotation nor prior knowledge. After estimating the model parameters from the patient record, we can use the resulting symptom and herb distributions for TCM clinical decision support. We outline three possible applications in the following subsections, though additional applications are also possible.

A. Mining Disease-Specific Symptoms and Herbs

Mining TCM symptoms and herbs can help doctors diagnose diseases as well as reveal when certain herbs are prescribed. We use a probabilistic model to provide distributions over symptoms and herbs, respectively. Distributions with the highest probabilities are the most relevant to disease \(d\).

B. Discovering Symptom-Herb Relationships

Besides mining symptoms and herbs of a given disease, our model can discover highly correlated symptoms and herbs. Here, we define the correlation between symptom \(s\) and herb \(h\) as:

\[
corr(s, h) = \sum_{d \in D} \Pr(s|d) \Pr(h|d)
\]

Thus, if symptom \(s\) and herb \(h\) co-occur frequently in many disease multinomial distributions, they will have a high correlation score. These symptom-herb relationships can help doctors understand the precise usages of each herb.

C. Disease and Treatment Prediction

Our model can be extended to build intelligent prediction systems that can predict a patient’s disease based on his or her symptoms. Assuming that all diseases are equally likely to occur, the probability of a patient with symptoms \(\{s_1, \ldots, s_n\}\) having disease \(d\) is computed as:

\[
\Pr(d|\{s_1, \ldots, s_n\}) \propto \Pr(s_1|d) \ldots \Pr(s_n|d) \Pr(d)
\]

Similarly, we can build an intelligent prediction system to recommend herbs to a patient based on his or her symptoms. The probability of a patient with symptoms \(\{s_1, \ldots, s_n\}\) receiving effective treatment by using herb \(h\) is computed as:

\[
\Pr(h|\{s_1, \ldots, s_n\}) \propto \sum_{d \in D_t} \Pr(s_1|d) \ldots \Pr(s_n|d) \Pr(h|d) \Pr(d)
\]

V. EXPERIMENTAL RESULTS

The main goal of our experiments is to examine whether the proposed model can indeed extract useful knowledge from patient records. While we envision that our method would be used to mine a large number of patient records from multiple physicians, direct experimentation with such a mixed dataset would make it difficult to choose the most suitable physician to assess the data mining results. This is because different physicians tend to have varying opinions due to the differences in their training and experiences. Thus, we decided to evaluate the proposed model using patient records from a single physician. This way, the physician can judge to what extent the mined results match the his reasoning and experience. Specifically, we used 10,907 anonymous electronic medical records provided by a TCM physician whose specialization is in digestive system treatments. We normalized the symptom and herb terms to obtain 9,530 records with 3,000 symptoms, 97 diseases, and 652 herbs. The most frequently occurring
disease is “chronic gastritis” and the most frequently occurring symptom is “abdominal pain and chills.”

We hypothesize that our model produces better results than the closely related, previously proposed symmetric generative model [13]. To test this, we explicitly model the asymmetric casual relations among symptoms, diseases, and herbs. Next, we assess the quality of our results by comparing them to those of the previous model. In addition, we also quantitatively evaluate the effectiveness of our model in predicting diseases and herbs based on actual diagnoses and prescriptions given by the physician. We now discuss our results in detail.

A. Quantitative Analysis of Disease Profiles

To quantitatively evaluate our model, we collected a set of herb-symptom relationship annotations, manually curated by TCM experts. These annotations contain 27,285 herb-symptom relationships. The symptoms and herbs of 1,624 of these relationships appear in our dataset. Consequently, we evaluate how our method can accurately identify this subset of 1,624 relationships.

We use AUROC, one of the most widely used metrics, to evaluate our method. We compare our method with frequent pattern mining (FPM), one of the most widely used approaches in pattern mining. Our method significantly outperforms FPM with p-value<4.45e-8 (Fig. 4). Numerically, our method achieves an AUROC of 0.7721, which is much higher than FPM’s AUROC of 0.6938. Notably, we found that FPM does not make accurate predictions when the false positive rate exceeds 0.2. This is because FPM requires an herb and a symptom to appear in the same transaction to constitute an herb-symptom relationship. Our model is not hampered by this restriction.

By associating a disease across different transactions, our model can identify herb-symptom relationships that never directly co-occur in any transaction. To further investigate the improvement from our method, we show the top ten herb-symptom relationships identified by our method but not by FPM (Table I). We can see that our method accurately identifies many meaningful herb-symptom relationships. All of these relationships were verified by TCM experts, and many of them are widely used in clinical practice.

B. Qualitative Analysis of Disease Profiles

The estimated model parameters Pr(s|d) and Pr(h|d) for disease d can be collectively regarded as a disease profile in the sense that they provide knowledge about the typical symptoms and herbs associated with a disease. We show three sample conditional symptom distributions with the highest entropy values among all 100 disease profiles learned with our model (Table II). The entropy of a distribution measures its information content. Thus, the higher the entropy, the more informative the distribution. Similarly, we show the herb distributions with the highest entropy values (Table III).

We show the top five symptoms, sorted by decreasing probability, for three different diseases (Table II). The physician judged these symptoms to be typical symptoms for the corresponding disease. Indeed, the relevance of these symptoms to the disease is often obvious. For example, “difficult bowel movements,” “abdominal distension,” and “dry, hard stools” are common symptoms for “constipation.” The fact that our model was able to learn these associations automatically from the dataset clearly demonstrates the effectiveness of the model for mining medical knowledge. Our model can also differentiate between similar diseases, such as “reflux esophagitis” and “chronic gastritis,” which share symptoms such as “acid reflux” and “heartburn.” We discover “paresthesia pharynges” and “discomfort in upper respiratory tract” for “reflux esophagitis,” while these symptoms do not appear for “chronic gastritis.”

Remarkably, many of these symptoms are unique TCM symptoms and are not well-studied in western medicine. For example, the model discovered that “dark, purplish tongue” is associated with “coronary heart disease.” Our physician verified these unique TCM symptoms for their corresponding diseases. This result further shows the effectiveness of using our model in revealing TCM knowledge from medical records. Our model can provide useful TCM symptoms to facilitate differential diagnosis and precision herb prescription.
correlated distributions may also help acquire TCM knowledge over symptoms, diseases, and herbs. While these general models simultaneously discovers three correlated distributions the previous model (Table IV). We observe that the previous model (Table IV) works better than the previous symmetric generative model [13], we show sample results from the same dataset using our model. As discussed earlier, it is crucial to consider the whole spectrum of symptoms of a particular individual to personalize his or her treatment. The results in Table V can be used to support for the treatment of future patients.

D. Disease and Treatment Prediction

We employ three standard metrics commonly used to evaluate prediction tasks: F1 score, precision, and recall. We take the macro-average of these three metrics (Table VI). We note that disease prediction based on symptoms achieves a reasonably high F1 score, precision, and recall (0.58, 0.73, and 0.61, respectively). Herb prediction, based on symptoms and diseases, achieves a relatively low F1 score, a much lower than the drug prediction results.
Chen et al. performed a comparative study of the five classification methods (e.g., support vector machines, neural networks, and decision trees) for syndrome differentiation in coronary heart disease using 1,069 clinical epidemiology survey cases [18]. The results showed that support vector machines (SVM) performed best in the prediction task. For the herb regularities, He et al. proposed an approach for discovering functional groups of herbs from a large set of prescriptions recorded in TCM texts [19]. Guang et al. constructed prescription association networks by mining literature datasets [20].

There are three studies of note most similar to our work [11, 13, 21]. Zhang et al. proposed the Symptom-Herb-Diagnosis topic modeling approach to discover the common relationships among symptom-herb combinations and disease diagnoses [11]. In another study, Zhang et al. proposed a topic model to automatically extract the hierarchical latent topic structures with both symptoms and herbs in TCM clinical data [13]. Although these works also use topic model-based approaches to mine TCM data, none of them explicitly model the asymmetric causal relations among symptoms, diseases, and herbs.

**VII. Conclusion**

We presented a new probabilistic model which captures the asymmetric causal relations of symptoms, diseases, and herbs in TCM patient records. Compared to previous models that analyze similar data, ours captures the asymmetry in the causal relations more accurately. The model discovers disease-symptom distributions and disease-herb distributions from medical records in a completely unsupervised manner.

An experienced physician judged our experimental results to be meaningful, suggesting that the proposed method is effective for discovering latent knowledge in TCM. Such knowledge can be very useful as references, especially to inexperienced doctors in clinical applications. The derived herb-symptom associations are also meaningful and provide useful insights and knowledge that can help personalize prescriptions.

Our current model is general and completely unsupervised, so it can be used to analyze large amounts of medical records collected by the TCM Data Center. This will be a major future direction for us to explore.
Another important future work is to perform natural language processing on the symptom descriptions to better understand the effectiveness of prescriptions. This information, coupled with the proposed model in this paper, will further enable us to distinguish groups of herbs that are effective for a disease from ineffective ones, directly advancing our understanding of treatments.

VIII. ACKNOWLEDGEMENTS

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 2013BAI02B01 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.

REFERENCES