Computer Systems Science & Engineering
DOI:10.32604/csse.2022.016759
Article

Tech Science Press

Mining Syndrome Differentiating Principles from Traditional Chinese Medicine Clinical Data

Jialin Ma[1,*], Zhaojun Wang[2], Hai Guo[3], Qian Xie[1,4], Tao Wang[4] and Bolun Chen[5]

[1]Jiangsu Internet of Things and Mobile Internet Technology Engineering Laboratory, Huaiyin Institute of Technology, Huaian, 223003, China
[2]Huaiyin Wu Jutong Institute of Traditional Chinese Medicine, Huaian, 223000, China
[3]The Affiliated Huaian No.1 People's Hospital of Nanjing Medical University, Huaian, 223000, China
[4]Jiangsu Eazytec Co., Ltd., Wuxi, China
[5]University of Fribourg, Fribourg, 1700, Switzerland
[*]Corresponding Author: Jialin Ma. Email: majl@hyit.edu.cn
Received: 11 January 2021; Accepted: 30 April 2021

Abstract: Syndrome differentiation-based treatment is one of the key characteristics of Traditional Chinese Medicine (TCM). The process of syndrome differentiation is difficult and challenging due to its complexity, diversity and vagueness. Analyzing syndrome principles from historical records of TCM using data mining (DM) technology has been of high interest in recent years. Nevertheless, in most relevant studies, existing DM algorithms have been simply developed for TCM mining, while the combination of TCM theories or its characteristics with DM algorithms has rarely been reported. This paper presents a novel Symptom-Syndrome Topic Model (*SSTM*), which is a supervised probabilistic topic model with three-tier Bayesian structure. In the *SSTM*, syndromes are considered as observed topic labels to distinguish certain symptoms from possible symptoms according to their different positions. The generation of our model is in full compliance with the syndrome differentiation theory of TCM. Experimental results show that the *SSTM* is more effective than other models for syndrome differentiating.

Keywords: TCM; syndrome differentiation; topic model; LDA; SSTM

1 Introduction

TCM has been existing for more than 3000 years. Different from modern orthodox biomedicine, it has an independently developed system of medical knowledge. The Therapy of TCM depends on natural herbs, acupuncture, scrape, cupping, etc. In addition, its diagnostic methods are unique via look, listen, question and feel the pulse. This is quite different from the modern medicine which mainly rely on medical instruments [1–3]. Furthermore, human body is considered as a synthetic system and abundant Chinese traditional naive philosophical thoughts are brought in TCM. Its aim is to adjust the ecological balance of human body rather than to treat an individual organ. Plenty of clinical practice has been proved that the TCM has unique effects on many special diseases like SARS, COVID-19, etc. In recent years, a growing number of foreigners have been accepting the treatment or health management of TCM [4–6].

One of the significant characteristic of TCM is to treat diseases based on syndrome differentiation. This is a process of comprehensive judgment based on analysis, induction, reasoning via four-ways information diagnosis. This is also the key link for doctors to select proper prescriptions or therapies [7–10]. However, the relationship between syndromes is complex. In clinical practice, there are several difficulties for syndrome differentiation as follows:

(a) Syndromes are complex, diverse and vague.

A disease is often accompanied with multiple syndromes, and different diseases share the same syndromes. Besides, the description of syndromes usually relies on natural language which is also polysemic and imprecise.

(b) The process of syndrome differentiation often has some subjectivity.

Syndrome differentiation is a process through which doctors make a diagnosis based on subjective knowledge and experience in accord with the objective reality of a patient. Because of the differences in individuals and the limited knowledge or experience of doctors, one patient may be diagnosed with different syndromes by different doctors [11–14].

In order to accurately master the complex structure of syndromes, and establish a diagnostic standard for TCM, in time, it is of great significance to analyze the principles of syndrome differentiation. This is beneficial for the inheritance, the improvement and the development of the diagnosis theory of TCM [15].

During the past thousand years of Chinese history, a huge number of TCM clinical data have been recorded as clinical cases, in ancient textbooks, and as classical ancient prescriptions etc. These data usually exist in the form of texts. At present, the technology of Text Mining (TM) has been frequently utilized to analyze TCM knowledge hidden in the large-scale textual data [5–9]. Feng et al. [12] provided an overview of knowledge discovery and KDD that had been applied in TCM. They emphasized that syndrome (Zheng) is a core element in TCM. They considered that KDD is a promising technology that can connect TCM and modern life sciences using TM and other knowledge discovery methods. Liu et al. [16] reviewed the application of TM in TCM knowledge discovery. They mentioned that using TM to analyzing TCM is still immature. They also pointed out that we great efforts should be made in acquiring richer semantic information in the future. In fact, syndrome differentiation (Bian Zheng) by the four ways information diagnosis is the first and the most important step in TCM clinical practice [7,8].

By observing massive TCM historical clinic records, we can find that usually more than one syndrome and a stack of symptoms were recorded for a patient. These syndromes can be considered as labels and symptoms as ‘word’ for these records. These textual records are very suitable to model by topic model which is a popular probabilistic graphical model. The topic model is an open generative modelling method that fully conforms to the Bayesian framework [17]. Although it has been successfully applied to latent semantic analysis and knowledge discovery, such as topic discovery, emotion analysis, and even image analysis, how to effectively combine the actual theory of analysis objects is the key [18,19]. Analysis syndrome differentiation principles began in Chinese ancient times. For example, in the famous TCM classic book “Shang Han Lun”, Zhongjing Zhang divided certain and possible symptoms in the process of syndrome differentiation.

In this paper, we present a generative probabilistic model—the Symptom-Syndrome Topic Model (*SSTM*) which is a three-level hierarchical Bayesian model. The *SSTM* can capture co-occurrence information between symptoms and syndromes from TCM history records. This study is a further expansion on the conference report we had published previously [20]. Compared with the previous version, the principle and parameters inference about *SSTM* are clarified more comprehensively in this paper. Besides, multiple experiments and comparisons are showed. In our *SSTM*, symptom-syndrome is modelled by a generative process, in which “Bian Zheng” (syndrome differentiation) knowledge or

principles from TCM historical records are acquired as patterns. Particularly, it is different from other relevant researches that we distinguish certain and possible symptoms in our model. This benefit to capture more effective latent relationship between symptoms and syndromes. Taken together, our method contributes to a better understanding of TCM diagnostic principles, and provides an effective model for computer automatic diagnosis.

In addition, literature review about TCM mining is summarized in the Section 2. The details and experimental results about *SSTM* are shown in the Sections 3 and 4 respectively. Finally, the conclusion and future work are illustrated in the Section 5.

2 Related Works

KDD has been a research hotspot in modern biomedicine field for many years, but its application in TCM has been highlighted in recent years [1–4,21–23]. Lukman et al. [24] surveyed the progress of computational approaches for TCM formulation and diagnosis mining. They considered that Bayesian networks (BNs) are available to capture relationships among complicated features from TCM records. However, topology learning for BNs is intractable due to the dramatically increased of features [25]. Miao et al. [15] focused on integrating syndrome differentiation with orthodox modern medical diagnosis in order to find novel methods for overall medical diagnosis and treatment [16]. In addition, a framework which can automatically mine treatment pattern from TCM clinical cases was proposed [26]. They introduced supervised topic models into field of TCM. Jiang et al. [13] directly applied the LinkLDA model to extract symptoms and their corresponding herbs. Zhao et al. [11] proposed a symptom-herb-diagnosis topic (SHDT) model to automatically extract the common relationships among symptoms, syndromes and formulas from large-scale TCM clinical data.

Recently, topic model has been becoming a hot theory in the research direction of TCM mining [20,27–28], which is one of a NB based on probability statistics theories. It can be used for detecting latent semantic structures and information in large-scale textual data [29–30]. In the early years, latent semantic analysis (LSA) as one of the famous representative method was used to capture word co-occurrence in documents [31]. The LSA can reveal the semantic dimensionality between texts and words. Then, probabilistic latent semantic analysis (PLSA) was progressed from LSA [32], in which a document is regarded as a mixture of topics, while a topic is a probability distribution over words. LDA was proposed for the first time by Blei in 2003 in order to improve the defects of the PLSA. Different from the PLSA, the LDA added Dirichlet priors in the distributions [17]. Therefore, LDA is a more completely generative model which had achieved great successes in TM and other artificial intelligence domains [33].

However, a standard LDA still cannot be directly used for TCM mining [34–36], because it is an unsupervised topic model, which unable to express observable syndromes. Another topic model labeled LDA is a probabilistic graphical model which can model multi-labeled document collection [19]. Nevertheless, labeled LDA fails to take into the theory about syndrome differentiation. Zhang et al. [11] captured a whole and general relationship among symptoms, herbs and diagnoses using SHDT, but they did not focus on the differentiating syndrome. This is a rare relevant study that mapped symptoms to syndromes and treatment methods. However, their method rely on the domain ontology base which usually is not easily to acquire [26,37]. The above-mentioned methods rarely focus on the relationship between syndromes and symptoms, they do not distinguish certain symptoms from possible ones which should be differentially treated in model. Therefore, our study intends to propose a novel model to identify the relationship between symptoms and syndromes based on relevant TCM theories. This is special for discovering syndrome principles from clinical textual data of TCM.

3 Our Work

This study is devoted to discover syndrome differentiation patterns from a large number of TCM clinical textual records by machine learning. Our work is beneficial to conclude the principles of "Bian Zheng". Although conventional probability topic models like LDA and labeled LDA are outstanding models for latent semantic analysis, relevant theory about syndromes differentiation is not considerate in these models. In this section, we present our *SSTM* and relevant TCM theory. The specific detail of the *SSTM* model and its parameters inferring are described in Sections 3.2 and 3.3 respectively. Finally, we present the framework of syndrome prediction based on *SSTM* in the Section 3.4.

3.1 Our Thinking and the Relevant TCM Theory

First of all, we observe the characteristics of TCM clinical records which come from China's national population and health science data sharing platform (http://www.ncmi.cn/). A part of three diabetes patients TCM clinical records show in Tab. 1.

Table 1: The records of diagnoses and symptoms about three diabetes patients

Patients	Syndromes	Symptoms
1	Deficiency of both Qi and Yin (气阴两伤证), internal obstruction of blood stasis (瘀血内阻证), deficiency of kidney yuan (肾元亏虚证)	Fatigue and weakness, lumbar debility, frequent urinate in night, edema, limb numbness, mouth parched and tongue scorched, vision decline
2	Damp-turbidity (湿浊证), deficiency of both Yin and Yang (阴阳两虚证), exogenous syndrome (外感证), damp heat syndrome (湿热证), blood stasis syndrome (血瘀证), deficiency of both Qi and Yin (气阴两虚证), deficiency of Yin of both the liver and kidney (肝肾阴虚证)	Numbness of hands and feet, ache, tongue dark, vein stasis under tongue, fatigue, weak, mouth parched and tongue scorched, lumbar debility, frequent urinate, emaciation, red tongue, moss thin and little jin, weak pulse, dizziness, tinnitus, vision decline, constipation
3	Yin deficiency and heat syndrome (阴虚燥热证), deficiency of both Qi and Yin (气阴两虚证), deficiency of the spleen and kidney (脾肾两虚证), deficiency of both Yin and Yang (阴阳两虚证)	Polydipsia, polyphagia, emaciation, red tongue, petechiae on the tip of the tongue, weak pulse, mouth parched and tongue scorched, thin, weakness, frequent urinate, frequent urinate in night, lumbar debility, limb edema, pale tongue, greasy tongue, mental weakness, chest tightness, nausea and retching

As shown in the Tab. 1, one patient often has multiple syndromes, and the relationships between syndromes and symptoms are diverse. These relationships have implicit structures, so their corresponding relationships are not explicit shown in these records. Furthermore, a patient usually has more than one syndrome. The symptoms in a record can be regarded as a document, and its corresponding syndromes are taken as multi-label. This relationships of syndromes-syndromes can be described in Fig. 1(a).

However, hidden relationships about important principle of individual differences in TCM records is neglected in Fig. 1(a). In TCM theory, the individual differences exist with changes of the climate, local and physical conditions. In fact, a patient usually has developed symptoms which mix up with individual symptoms. Therefore, syndromes of each patient vary a lot, and individualized syndromes also highlight the importance of the individualized diagnosis and treatment in TCM. Some symptoms are not useful for

diagnosis, but they may even interfere the syndrome differentiation. One syndrome usually has some basic and individual symptoms, which was also mentioned in many TCM literatures, such as “Shang Han Lun” (Febrile Diseases). This principle was followed by TCM exporters who were even organized by China Administration to develop standards for syndromes. Therefore, according to the relevant theory of “Bian Zheng”, symptoms are divided into two different subtypes in our model.

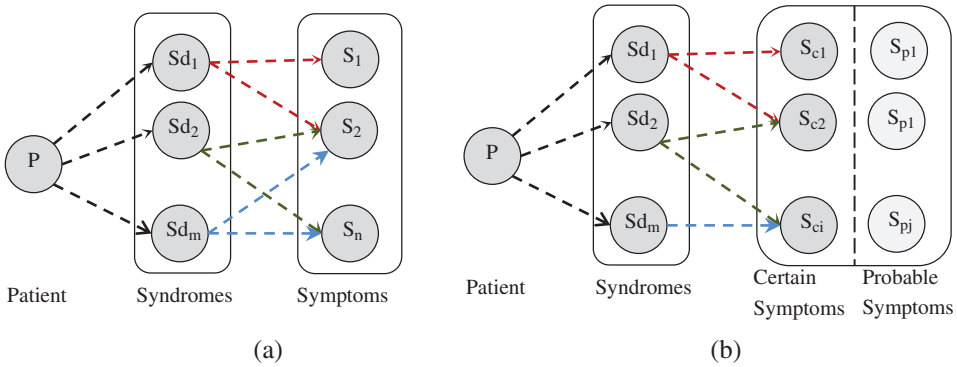

Figure 1: The relationships between symptoms and syndromes, (a) general and (b) comprehensive

• **Certain symptoms** are some main or necessary symptoms for a syndrome. They often determine pathogeny, location and nature of a disease.

• **Possible symptoms** are some uncertain symptoms for a syndrome. They often possibly appear, and unimportant or subordinate for diagnosis, but they can be transformed into primary symptoms sometimes.

According to the above definition, a more comprehensive relationships of symptoms-symptoms for a patient is shown in Fig. 1(b). The difference between certain symptoms and possible symptoms is specifically concerned in our model, which benefits to effectively capture latent relationships between symptoms and syndromes.

3.2 Our SSTM

Conventional topic models, like PLSA [32], LDA [17], supervised LDA [30], and labeled LDA [19] etc., reveal the latent topics within *corpus* by implicitly capturing the document-level word co-occurrence patterns. Despite all of this, all terms in these standard topic models are considered as the same status [38]. However, such equal status of terms is not suitable to some fields where terms play different roles on the semantic contribution. This phenomenon is particularly evident for syndromes.

We propose a generative probabilistic topic model *SSTM* to capture the relationship between symptoms and the syndromes from TCM records. It is a three-level hierarchical Bayesian model, in which symptoms are distributed in a syndrome, and observable variable symptoms are distinguished into certain and possible symptoms. This distinguishing type of symptoms complies with the theory of TCM, which is more beneficial to capture effective co-occurrence information for syndrome differentiating.

In the *SSTM*, TCM clinical records set is considered as the number of D independent text generation process. The symptoms in a record is regard as a ‘document’ and a symptom is a ‘term’. Syndromes are taken as observed topic labels corresponding to a series of symptoms. Syndromes can be seen as a multi-topic labels set for the ‘document’. Similar to standard topic model, document d is composed of the number of K topics in proportion, which satisfies multinomial distribution in *SSTM*. The number of K topics is also composed of words (symptoms), which also satisfies multinomial distribution. According to Bayesian theory, all parameters need to satisfy a certain distribution. Here, Dirichlet distribution is the

prior for multinomial distribution. Different from the classical LDA model, the *SSTM* is a supervised model, in which syndromes are observed constraint in the generating of topics. Consistent with the labeled LDA, the *SSTM* focuses on the distinction between the certain and possible symptoms for a patient, which is controlled by a binary variable y who satisfies Bernoulli distribution, and its prior is Beta distribution. Therefore, the *SSTM* conforms relevant theory of TCM, and it is conductive to capture syndrome differentiating patterns.

The graphical model of the *SSTM* is shown in Fig. 2. All symbols are annotated in Tab. 2. Vector $L = \{1,2,3\ldots,K\}$ represents a set of the topics labels (syndromes), and vector $\Lambda_{ds} = (t_1, t_2,\ldots, t_K)$ is a list of binary topic presence/absence indicator of a document d ($d \in D$), where each $t_k \in \{0,1\}$. Let φ_k denotes the symptom distribution on topics and ψ_d denotes the subordinate symptoms distribution. Λ_d obeys Bernoulli distribution, in which θ_d is only restricted over the topics corresponding to Λ_d. λ_d denotes Bernoulli distribution which controls the indicator y_{dn} for choice between certain or possible symptom. φ_k, θ_d, and ψ_d all obey multinomial distributions. Their prior distributions are drawn from the symmetric Dirichlet (β), Dirichlet (α), and Dirichlet (η) respectively. Moreover, λ_d and Λ_d are drawn from prior distributions Beta (γ) and Beta (ν) respectively. The probability of a symptom s_{dn} is formulated as follows:

$$p(s_{dn}|d) = p(y_{dn} = 0|d) \sum_z p(z_{dn}|d)p(s_{dn}|z_{dn}) + p(y_{dn} = 1|d)p(s_{dn}|y_{dn} = 1) \quad (1)$$

The generation processes of clinical data set by *SSTM* are listed as follows:

Step 1: For each topic $k = 1,2,\ldots,K$:

Draw $\varphi_k \sim Dir\ (\beta)$

Step 2: For each document $d = 1,2,\ldots D$:

Step 2-1; Draw $\psi_d \sim Dir\ (\eta)$, Draw $\lambda_d \sim Beta\ (\nu)$,

Step 2-2; For each topic $k \in \{1,\ldots,K\}$

Draw $\Lambda_{dk} \in \{0,1\} \sim Beta\ (\gamma)$

Step 2-3; Draw $\theta_d \sim Dir(\alpha|\Lambda_{dk})$, where $t_i = 1$

Step 2-4; For each symptom $s = 1,2,\ldots N_d$

(I) Draw $y \sim Bernoulli\ (\lambda_d)$

(Π) if $y = 1$ then

Draw $s_{dn} \sim Multi\ (\psi_d)$

Else

Draw $z_{dn} \sim Multi\ (\theta_d)$,

Draw $s_{dn}|\ z_{dn}, \varphi_k \sim$Multinomial (φ_{zdn})

3.3 Parameters Inference

As a probability graph model, the *SSTM* contains several observed and latent variables. As shown in Fig. 2, grey node variables $\boldsymbol{\Lambda}$ (syndrome label) and $\boldsymbol{S}$ (symptom) are observed variables, the hollow node variables $\boldsymbol{y}$, $\boldsymbol{z}$, $\boldsymbol{\lambda}$, $\boldsymbol{\theta}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ are all unknown latent variables requiring to be inferred. The super parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\nu}$, $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ can be manually set according to experiences or experiments. Therefore, they can be regarded as known variables. In *SSTM*, certain or possible inferring for each symptom in document ($y = 0$ or $y = 1$) is the first key event. This follows by the inference of the syndrome label z for the certain symptom. Finally, other parameters $\boldsymbol{\lambda}$, $\boldsymbol{\theta}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ can be calculated indirectly.

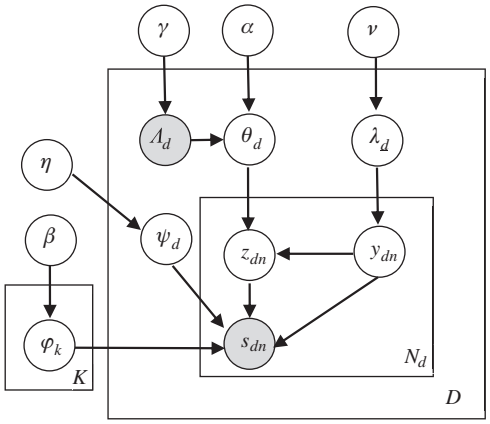

Figure 2: The graphical models of the *SSTM*

Table 2: The annotation of symbols for the *SSTM*

Symbols	Annotation
D	The total number of dataset.
K	The total number of topics.
N_d	The total number of symptoms in document d.
s_{dn}	A symptom in d.
z_{dn}	A topic number appearing in d.
φ_k	A multinomial distribution denotes symptoms distribution on topic k.
ψ_d	A multinomial distribution denotes distribution of possible symptoms in d.
θ_d	A multinomial distribution denotes syndromes distribution on document d.
λ_d	A Beta (ν) distribution.
Λ_d	A Beta (γ) for explicitly constraining θ_d.
y_{dn}	A Bernoulli distribution.
β	Hyper parameter for φ_k.
η	Hyper parameter for ψ_d.
α	Hyper parameter for θ_d.

Multiple methods can be used for parameter estimation of a topic model, such as expectation maximization algorithm (EM), variation inference, maximum *a posteriori* estimation and Gibbs sampling. All of them estimate parameters approximately, because accurate inference is difficult to achieve. Gibbs sampling is an effectively and widely used Markov chain Monte Carlo algorithm for latent variable inference [27,33]. Therefore, we adopt Gibbs sampling to estimate the latent variables in the *SSTM*. Here, a collapse Gibbs sample method is used to infer the latent variables y_{dn} and z_{dn}. The specific formulas deduction process are as follows:

$$
\begin{aligned}
&p(y_{dn}=1|\mathbf{y}_{-dn},\mathbf{S},\mathbf{z}_{-dn}) \propto p(y_{dn}=1,s_{dn}=t|\mathbf{y}_{-dn},\mathbf{S}_{-dn},\mathbf{z}_{-dn})\\
&=\int p(y_{dn}=1,s_{dn}=t,\lambda_d,\psi|\mathbf{y}_{-dn},\mathbf{S}_{-dn},\mathbf{z}_{-dn})d\lambda_d d\psi\\
&=\int p(\lambda_d|y_{-dn})p(y_{dn}=1|\lambda_d)p(\psi|\mathbf{S}_{-dn})p(\mathbf{S}_{dn}=t|\psi)d\lambda_d d\psi\\
&=\int (p(\lambda_d|y_{-dn})p(y_{dn}=1|\lambda_d)d\lambda_d \int (p(\psi|\mathbf{S}_{-dn})p(s_{dn}=t|\psi)d\psi\\
&=\frac{n^{(d)}_{-dn,y=1}+\upsilon}{n^{(d)}_{-dn,y=\cdot}+2\upsilon}\frac{n^{(y=1)}_{-dn,t}+\eta}{n^{(y=1)}_{-dn,\cdot}+V\eta}
\end{aligned}
\tag{2}
$$

$$
\begin{aligned}
&p(y_{dn}=0,z_{dn}=k|\mathbf{y}_{-dn},\mathbf{S},\mathbf{z}_{-dn}) \propto p(y_{dn}=0,z_{dn}=k,s_{dn}=t|\mathbf{y}_{-dn},\mathbf{S}_{-dn},\mathbf{z}_{-dn})\\
&=\int p(y_{dn}=0,z_{dn}=k,s_{dn}=t,\lambda_d,\theta_d,\varphi_k|\mathbf{y}_{-dn},\mathbf{S}_{-dn},\mathbf{z}_{-dn})d\lambda_d d\theta_d d\varphi_k\\
&=\int p(\lambda_d|y_{-dn})p(y_{dn}=1|\lambda_d)p(\varphi_k|\mathbf{S}_{-dn})p(s_{dn}=t|\varphi)p(\theta_d|\mathbf{z}_{-dn})d\lambda_d d\theta_d d\varphi_k\\
&=\frac{n^{(d)}_{-dn,y=0}+\upsilon}{n^{(d)}_{-dn,y=\cdot}+2\upsilon}\frac{n^{(k)}_{-dn,t}+\beta}{n^{(k)}_{-dn,\cdot}+V\beta}\frac{n^{(d)}_{-dn,k}+\alpha}{n^{(d)}_{-dn,\cdot}+M_d\alpha}
\end{aligned}
\tag{3}
$$

In Eqs. (2) and (3) '–' refers to exclude the current instance; '.' represents all; $n^{(d)}_{-dn,y=1}$ represents the count of all symptoms with $y = 1$ after excluding the current position symptom in d; $n^{(y=1)}_{-dn,t}$ represents the count of symptom t marked $y = 1$ in the whole data set when the current instance is excluded; M_d represents the number of syndrome in document d; V represents the number of symptoms. Other symbols have similar meanings. The above formulas are two Gibbs sampling for latent variables y and z. In addition, according to Bayesian principle, other latent parameters $\boldsymbol{\lambda}$, $\boldsymbol{\theta}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ can be obtained by the following posterior estimation using average method:

$$\hat{\theta}_{dk}=\frac{n^{(d)}_{-dn,k}+\alpha}{n^{(d)}_{-dn,\cdot}+M_d\alpha} \tag{4}$$

$$\hat{\varphi}_{kt}=\frac{n^{(k)}_{-dn,t}+\beta}{n^{(k)}_{-dn,\cdot}+V\beta} \tag{5}$$

$$\hat{\lambda}_d=\frac{n^{(d)}_{-dn,y=0}+\upsilon}{n^{(d)}_{-dn,y=\cdot}+2\upsilon} \tag{6}$$

$$\hat{\psi}_d=\frac{n^{(y=1)}_{-dn,t}+\eta}{n^{(y=1)}_{-dn,\cdot}+V\eta} \tag{7}$$

3.4 Syndrome Prediction

Automatic learning syndromes differentiation patterns from TCM clinical records by computers has been settled by the proposed *SSTM* above. How to predict syndromes in the condition with unlabeled symptoms for a patient. It means we need to infer syndromes (topics or labels) for unlabeled records by

the trained *SSTM*. According to the four ways of diagnosis information, automatic diagnosis should be realized by computers.

According to Bayesian theory, the prior and posterior parameters in *SSTM* follow the same distribution. They are conjugate distributions which can be used to predict syndromes labels when give new observed variables. The model parameters of prior distribution have been solved by Gibbs sampling in the section 3.3. Because the *SSTM* is a supervised model, it cannot be directly used for syndrome prediction. To solve this problem, we relax the *SSTM* by removing the constraint Λ_d. Fig. 3 shows the framework of syndrome prediction, which is based on the *SSTM*.

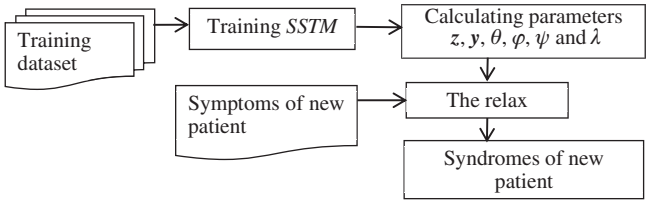

Figure 3: Framework for syndromes prediction based on the *SSTM*

4 Experiment

In our study, two kind of experiments were conducted for models assessing. The first kind were to evaluate objective performance indexes, including perplexity, Kullback-Leibler and time consuming between the *SSTM* and other traditional probabilistic topic models (LDA and labeled LDA). The second kind experiments were performed to compare model learning and prediction effectiveness.

4.1 Dataset

Experimental data were obtained from the Chinese traditional medicine database of national population and health science data sharing platform (http://www.ncmi.cn/index.html). This dataset contains 24 subject databases with a total of 1,741 diseases, and involves 127,541 pieces of records. After removing records which lack syndromes or symptoms, a total of 8,653 pieces of eligible records were used for our experiments. The statistical information of the data set is shown in Tab. 3

Table 3: Basic information statistics for experimental dataset

Symbol	Syndromes	Symptoms
Total frequency	6,444	33,360
Dictionary size	166	5,572
Average quantity	3.7 syndromes per record	19.2 symptoms per record

4.2 Performance Evaluation and Analysis

The indexes perplexity and Kullback Leibler (KL) distance are commonly used for performance evaluation of probabilistic topic models. Perplexity is used to evaluate prediction ability by test data, which reflects the ability of generalization. Perplexity is negatively correlated to the prediction ability; the lower perplexity indicates the stronger prediction ability of the model. Here, the perplexity is used to evaluate the performance of LDA and CTM [38]. Its calculation formula is listed as follows:

$$Perp(\Phi) = \left(\sum_{d=1}^{D}\sum_{i=n+1}^{N_d} p(w_i|\Phi, w_{1:n})\right)^{\frac{-1}{\sum_{d=1}^{D}(N_d - n)}} \tag{8}$$

In the Eq. (8), Φ represents parameters of the evaluated model.

KL distance is usually used to test the distinguishing ability of the model by calculating the average similarity between topics. The lower similarity indicates a higher quality of the topics. Because of the asymmetry, the symmetrical KL (sKL) divergence is often used to calculate the average distance between topics [33]. The specific formula of sKL is as follows:

$$sKL(\varphi_i, \varphi_j) = \sum_{k=1}^{K}\left(\varphi_{ik}\log\frac{\varphi_{ik}}{\varphi_{jk}} + \varphi_{jk}\log\frac{\varphi_{jk}}{\varphi_{ik}}\right) \tag{9}$$

In Eq. (9), K represents the total number of topics; φ_i and φ_j represent the probability distribution of two topics words.

In our experiment, 90% of the experimental records were randomly selected for model training, and the remaining were used to test. We calculated the perplexity and sKL distance of the LDA, labeled and *SSTM* respectively. The common parameters of the three models referred to a previous study, where $\alpha = 0.01$, and $\beta = K/50$ [33]. $K = 166$ is the number of unique syndromes in the dataset. The super parameter $\eta = 0.01$ and $\alpha = 0.5$ were set for *SSTM*. The iterations for the three models were set to inter = 1000. The experiments were run on a computer with 8 GB memory and dual Core Intel processor i5-6500. The indexes of perplexity, sKL, and time consuming among *SSTM*, LDA and labeled LDA are shown in Figs. 4–6.

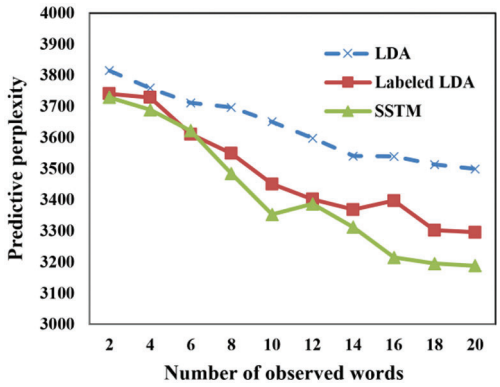

Figure 4: The perplexity indexes of the three models

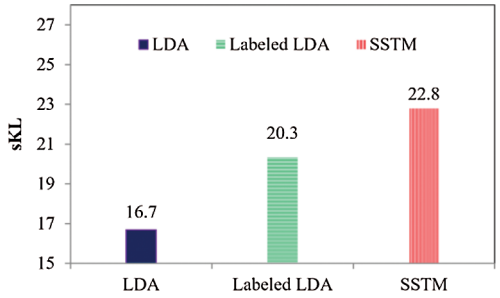

Figure 5: The sKL indexes of the three models

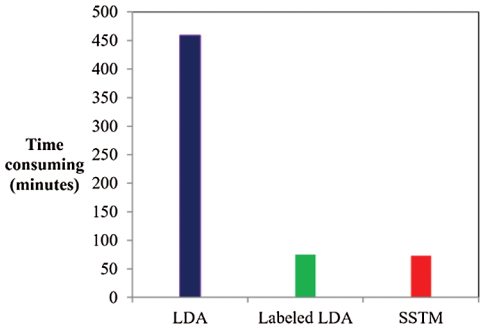

Figure 6: Time-consuming of the three models

As shown in Fig. 4, with the increasing number of observed symptoms in the test data (from 2 to 20 steps 2), the perplexity indexes of the three topic models were calculated. All the perplexity points of LDA are almost higher than the other two models (Fig. 4). The *SSTM* is lower than or close to the labeled LDA at most of time. It means that the prediction ability of the *SSTM* is better than labeled LDA. The *SSTM* and labeled LDA both performed better than LDA. Because we take the interference of possible symptoms for syndrome differentiation into consideration in *SSTM*, its prediction ability is better than labeled LDA. Fig. 5 shows the average sKL divergence of the *SSTM* is also higher than the other two models. We set $K = 166$ for LDA model and the other two supervised models, but syndrome assigning is an unsupervised process. Therefore, the prediction and quality of LDA are more inferior than other two models.

The experimental program was written in 'Pycharm 2017' with Python 3.5. According to the principle of the LDA, labeled LDA and *SSTM*, all of their algorithm time complexity are $\boldsymbol{O}$ (*Inter* * *M* * *N*), where *Inter* represents iterations, *M* is the total number of documents, and *N* is the total number of words in the dataset. Fig. 6 shows the consumption of training time of the three models. Obviously, the training time of LDA remains the longest, which is about 7 times as much as that of the other two models. It may be attributed to the calculation of posteriori probability for each symptom of all syndromes ($k = 166$) in the training process. In contrast, labeled LDA and *SSTM* only calculated the probability of the syndromes in the explicit constraint set Λ_d, which greatly reduce the amount of calculation and the time consumption.

4.3 Effectiveness Evaluation and Analysis

Fig. 7 shows some representative visual cases of the *SSTM* training results. In addition, 10 syndrome examples learned by the *SSTM* are depicted in Tab. 5, where the top 6 symptoms and their probabilities of the example are listed. Furthermore, the definition of the above 10 syndromes in Baidu Encyclopedia knowledge base is taken as a baseline for comparing the effect of learning syndrome differentiation patterns between *SSTM* and labeled LDA. The accuracy of comparison results is calculated by the follows formula:

$$Precision(sd) = \frac{count(\{s_0, s_1, \ldots s_{N_{top,sd}}\} \cap \{s_0, s_1, \ldots s_{N_{Baidu,sd}}\})}{N_{top,sd}} \tag{10}$$

In Eq. (10), *sd* represents a syndrome; *s* represents a symptom; $N_{top,sd}$ represents the number of top *N* symptoms selects according to the probability value about *sd*. $N_{Baidu,sd}$ represents the number of symptoms listed in the definition of Baidu Encyclopedia knowledge about *sd*; *count*() denotes the number of intersection symptoms of two sets. Tab. 4 shows the precision of learning for the 10 syndromes differentiation patterns by *SSTM* and labeled LDA.

Syndromes: Yin deficiency of liver and kidney, deficiency of lung Qi, Yin deficiency and lung dryness, deficiency of lung Qi and cold
Symptoms: Coughing, mental exhaustion, emaciation, short of breath, chest tightness, asthma, spontaneous sweating, tongue fur thin white, expectoration, cold and limb cold, five heart dysphoria, thready rapid pulse, less fur, weak waist and knees, bloody sputum, pharynx trunk, pale or dull complexion, fat and thin tongue, pale or purplish tongue, hot flashes and night sweats, dry cough and less sputum, pulse sinking, cold limbs, wheezing, hoarseness, chills, faint voice, red tongue tip, late pulse, night sweat.
*__Note:__ Syndrome and its certain symptoms were expressed as the same font color; Probable symptoms were expressed as the black color font.

Figure 7: A representative visual case of *SSTM* training results

Table 4: The precision of 10 syndromes patterns learning by *SSTM* and labeled LDA

No.	Syndromes (Topics)	Labeled LDA	*SSTM*
1	Deficiency Yang of spleen and kidney	0.62	0.75
2	Deficiency of Qi and Yin	0.45	0.68
3	Liver depression and spleen deficiency	0.67	0.78
4	Qi stagnation and blood stasis	0.29	0.44
5	Damp-heat syndrome	0.64	0.74
6	Yin deficiency syndrome	0.31	0.55
7	Qi deficiency and blood stasis	0.48	0.59
8	Phlegm stasis and heat	0.67	0.76
9	Qi deficiency syndrome	0.74	0.85
10	Asthenia of spleen and stomach	0.66	0.72

Table 5: Ten examples of syndromes differentiating patterns acquired by the SSTM

No.	Syndromes (Topics)	Top 6 symptoms and their probability values
1	Deficiency Yang of spleen and kidney	Deep and thin pulse (0.0597), mental exhaustion (0.0523), pale tongue (0.0470), wheezing (0.0453), spontaneous sweating (0.0436), tinnitus (0.0432)
2	Deficiency of Qi and Yin	Limb cold (0.0596), chilly (0.0498), tongue fat (0.0471), emaciation: 0.0437, mental fatigue (0.0419), dry cough and less phlegm (0.0417)
3	Liver depression and spleen deficiency	Anaerobic greasy (0.0520), thin tongue coating (0.0517), reddish tongue (0.0449), belching noise (0.0405), low heat and night sweat (0.0383), low heat fluctuation (0.0380)
4	Qi stagnation and blood stasis	Mouth dry and bitter (0.0148), soreness of waist (0.0118), slightly greasy tongue coating (0.0109), teeth mark on the sharp edge (0.0108), not warm or swollen limbs (0.0107), anorexia (0.0107)
5	Damp-heat syndrome	Feeling plug throat (0.0470), belching noisy (0.0347), cold (0.0276), powerful pulse (0.0273), low heat and night sweat (0.0270), nasal flaring (0.0260)

Table 5 (continued).

No.	Syndromes (Topics)	Top 6 symptoms and their probability values
6	Yin deficiency syndrome	Chilly limbs (0.0283), low heat and night sweat (0.0248), loose stools (0.0207), abdominal cold and pain (0.0205), sweat like beads (0.020014), epigastria pain (0.019918)
7	Qi deficiency and blood stasis	Thin tongue coating (0.0212), poor cough (0.0162), excessive phlegm (0.0161), greasy mouth (0.0161), cough and breath (0.0147), chest tightness and shortness of breath (0.0139)
8	Phlegm stasis and heat	Scorching and pricking (0.0359), chest tightness (0.0317), yellow phlegm (0.0306), dark red tongue (0.0298), yellow and greasy of tongue coating (0.0203), astringent pulse (0.0197)
9	Qi deficiency syndrome	Pharyngeal itch (0.0247), fat tongue (0.0189), harsh language (0.017), light tongue (0.0171), ecchymosis on the edge of tongue (0.0162), dry eyes (0.0162)
10	Asthenia of spleen and stomach	Abdominal colic (0.0406), belching noise (0.0376), pulse strength (0.0328), limb tenderness (0.0292), acid regurgitation (0.0288), thin tongue coating (0.0250)

Tab. 5 shows the comparison of the precision between the labeled LDA and *SSTM* for learning patterns of syndromes differentiation. The manual evaluation was selected to define syndrome terms in Baidu Encyclopedia as the baseline. Synonyms were manually treated in the process of comparison. We selected the top 8 symptoms with the highest probability learned by the two models as the comparison objects. As shown in Tab. 4, the precision of the *SSTM* for all syndromes of 10 is 10%–20% higher than labeled LDA. The above results demonstrate that the *SSTM* is more effective than labeled LDA in the task of syndromes differentiation, mainly because the interference of possible symptoms in syndromes differentiation is taken into consideration in the *SSTM*. Therefore, the *SSTM* is capable of capturing purer syndrome patterns.

5 Conclusion and Future Work

We present a novel Symptom-Syndrome Topic Model (*SSTM*) that can effectively analyze complex and changeable syndrome differentiation patterns from TCM historical clinic records. These acquired patterns reify abundant diagnosis experiences and principles from TCM doctors. Furthermore, we then present a computer automatic diagnosis method (syndromes differentiating) based on the *SSTM*. The *SSTM* is characterized by the combination of relevant theories of TCM. We consider the interference of possible symptoms for syndrome differentiation in our model, which is more consistent with diverse clinical manifestations due to individual differences. Our experimental results confirmed the better quality of syndrome differentiation patterns learned by *SSTM* than labeled LDA. This study provides a method for TCM intelligent diagnosis. However, this novel model requires annotated data sets which are often difficult to obtain. The future work should be extended to semi-supervised or unsupervised learning model.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Z. Liu, Z. Y. Zheng, X. W. Guo, L. Qi, J. Gui *et al.,* "Attentive herb: A novel method for traditional medicine prescription generation," *IEEE Access*, vol. 7, no. 59, pp. 139069–139085, 2019.

[2] A. Mirarab, S. L. Mirtaheri and S. A. Asghari, "A model to create organizational value with big data analytics," *Computer Systems Science and Engineering*, vol. 35, no. 2, pp. 69–79, 2020.

[3] G. Zhang, Y. Huang and X. Zhang, "Deep feature learning based clustering with application to TCM data analysis," in *Proc. ITME*, Beijing, BJ, China, pp. 750–754, 2018.

[4] L. Yao, Y. Zhang, B. G. Wei, W. J. Zhang and Z. Jin, "A topic modeling approach for traditional Chinese medicine prescriptions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1007–1021, 2018.

[5] Y. H. Gu, Y. Wang, C. L. Ji, P. Fan, T. Wang *et al.,* "Syndrome differentiation of IgA nephropathy based on clinic pathological parameters: A decision tree model," *Evidence-Based Complementary and Alternative Medicine*, vol. 2017, no. 2, pp. 1–11, 2017.

[6] Y. F. Zhao, L. Y. He, G. Z. Li, B. Y. Liu, J. Wang *et al.,* "A novel classification method for syndrome differentiation of patients with AIDS," *Evidence-Based Complementary and Alternative Medicine*, vol. 2015, no. 6, pp. 1–8, 2015.

[7] L. Yao, Y. Zhang, B. G. Wei and Z. Jin, "Incorporating knowledge graph embedding into topic modeling," in *Proc. 31th AAAI Conf. on Artificial Intelligence*, San Francisco, USA, pp. 3119–3126, 2017.

[8] X. L. Zhou, Y. G. Liu, Q. Q. Li, Y. Zhang and C. B. Wen, "Mining effective patterns of Chinese medicinal formulae using top-k weighted association rules for the Internet of medical things," *IEEE Access*, vol. 6, no. 11, pp. 57840–57855, 2018.

[9] M. Chen, Y. Hao, K. H. Wang and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, no. 15, pp. 8869–8879, 2017.

[10] J. Chen, D. X. Yang, Y. Cao and Y. Y. Ma, "Syndrome differentiation and treatment algorithm model in traditional Chinese medicine based on disease cause, location, characteristics and conditions," *IEEE Access*, vol. 6, no. 21, pp. 71801–71813, 2018.

[11] X. P. Zhao, X. Z. Zhou, H. K. Huang, Q. Feng and B. S. Chen, "Topic model for Chinese medicine diagnosis and prescription regularities analysis: Case on diabetes," *Chinese Journal of Integrative Medicine*, vol. 17, no. 4, pp. 307–313, 2011.

[12] Y. Feng, Z. H. Wu, X. Z. Zhou, Z. M. Zhou and W. Y. Fan, "Knowledge discovery in traditional Chinese medicine: State of the art and perspectives," *Artificial Intelligence in Medicine*, vol. 38, no. 3, pp. 219–236, 2006.

[13] Z. Jiang, X. Zhou, X. Zhang and S. Chen, "Using link topic model to analyze traditional Chinese medicine clinical symptom-herb regularities," in *Proc. IEEE 14th Int. Conf. on e-Health Networking, Applications and Services*, Beijing, BJ, China, pp. 15–18, 2012.

[14] Z. Huang, W. Dong, L. Ji, C. H. He and H. L. Duan, "Incorporating comorbidities into latent treatment pattern mining for clinical pathways," *Journal of Biomedical Informatics*, vol. 59, no. 54, pp. 227–239, 2016.

[15] J. Miao, C. Lu, C. Zhang, J. Yang, Y. Tan *et al.,* "Syndrome differentiation in modern research of traditional Chinese medicine," *Journal of Ethnopharmacology*, vol. 140, no. 3, pp. 634–642, 2012.

[16] B. Y. Liu, X. Z. Zhou, Y. H. Wang, Q. J. Hu, L. Y. He *et al.,* "Data processing and analysis in real-world traditional Chinese medicine clinical data: Challenges and approaches," *Statistics in Medicine*, vol. 31, no. 7, pp. 653–660, 2012.

[17] D. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 2, pp. 993–1022, 2003.

[18] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[19] D. Ramage, D. Hall, R. Nallapati and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing*, SG, Singapore, pp. 248–256, 2009.

[20] J. L. Ma and Z. J. Wang, "Discovering syndrome regularities in traditional Chinese medicine clinical by topic model," in *Proc. Int. Conf. on P2P, Parallel, Grid, Cloud and Internet Computing*, Asan, Korea, pp. 157–162, 2016.

[21] J. Ma, J. Cheng, L. Zhang, L. Zhou and B. Chen, "A phrase topic model based on distributed representation," *Computers Materials & Continua*, vol. 64, no. 1, pp. 455–469, 2020.

[22] P. Kalaivaani and D. R. Thangarajan, "Enhancing the classification accuracy in sentiment analysis with computational intelligence using joint sentiment topic detection with medlda," *Intelligent Automation & Soft Computing*, vol. 26, no. 1, pp. 71–79, 2020.

[23] Z. Zhou, J. Qin, X. Xiang, Y. Tan, Q. Liu *et al.,* "News text topic clustering optimized method based on TF-IDF algorithm on spark," *Computers Materials & Continua*, vol. 62, no. 1, pp. 217–231, 2020.

[24] S. Lukman, Y. L. He and S. C. Hui, "Computational methods for Traditional Chinese medicine: A survey," *Computer Methods and Programs in Biomedicine*, vol. 88, no. 3, pp. 283–294, 2007.

[25] Z. Wu, H. Chen and X. Jiang, "Overview of knowledge discovery in traditional Chinese medicine," *Modern Computational Approaches to Traditional Chinese Medicine*, vol. 32, no. 10, pp. 1–26, 2012.

[26] L. Yao, Y. Zhang, B. Wei, W. Wang, Y. Zhang *et al.,* "Discovering treatment pattern in traditional Chinese medicine clinical cases using topic model and domain knowledge," in *Proc. 2014 IEEE Int. Conf. on Bioinformatics and Biomedicine*, London, UK, pp. 191–192, 2014.

[27] C. Chemudugunta, P. Smyth and M. Steyvers, "Modeling general and specific aspects of documents with a probabilistic topic model," *Advances in Neural Information Processing Systems*, vol. 13, no. 5, pp. 241–248, 2007.

[28] Z. Huang, X. Lu and X. Duan, "Latent treatment pattern discovery for clinical processes," *Journal of Medical Systems*, vol. 37, no. 2, pp. 9915, 2013.

[29] X. Yan, J. Guo, Y. Lan and X. Cheng, "A biterm topic model for short texts," in *Proc. of the 22nd Int. Conf. on World Wide Web*, Brazil: Rio de Janeiro, pp. 1445–1156, 2013.

[30] D. M. Blei and J. D. Mcauliffe, "Supervised topic models," *Advances in Neural Information Processing Systems*, vol. 20, pp. 121–128, 2010.

[31] K. Thomas, P. W. F. Landauer and L. Darrell, "An introduction to latent semantic analysis," *Discourse Processes*, vol. 25, no. 259, pp. 259–284, 1998.

[32] T. Hofmann, "Probabilistic latent semantic indexing," *Proc. of the 22nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Berkeley, California, USA, vol. 41, no. 6, pp. 50–57, 1999.

[33] G. Heinrich, "Parameter estimation for text analysis," Technical Report, 2005.

[34] N. Esfandiari, M. R. Babavalian, A. M. E. Moghadam and V. K. Tabar, "Knowledge discovery in medicine: Current issue and future trend," *Expert Systems with Applications*, vol. 41, no. 9, pp. 4434–4463, 2014.

[35] X. W. Wang, H. B. Qu, P. Liu and Y. Y. Cheng, "A self-learning expert system for diagnosis in traditional Chinese medicine," *Expert Systems with Applications*, vol. 26, no. 4, pp. 557–566, 2004.

[36] X. L. Xie, C. J. Lu, Z. Zeng and Z. P. Zeng, "Research on the psoriasis vulgaris syndrome differentiation standard of traditional Chinese medicine based on data mining technology," in *Proc. BIBM*, Shanghai, SH, China, pp. 281–284, 2013.

[37] L. Yao, Y. Zhang, B. G. Wei, W. Wang, Y. J. Zhang *et al.,* "Discovering treatment pattern in traditional Chinese medicine clinical cases by exploiting supervised topic model and domain knowledge," *Journal of Biomedical Informatics*, vol. 58, no. 29, pp. 260–267, 2015.

[38] D. Blei and J. Lafferty, "Correlated topic models," in *Proc. of the 18th Int. Conf. on Neural Information Processing Systems*, Cambridge, United States, pp. 147–154, 2005.