Disease Topic Modeling of Users' Inquiry Texts: A Text Mining-Based PQDR-LDA Model for Analyzing the Online Medical Records

Xin Liu, Yanju Zhou, Zongrun Wang, Ajay Kumar, and Baidyanath Biswas

***Abstract*—Disease information mining is one of the critical factors affecting users' perception of the disease and has attracted extensive attention from the information management community in recent years. If the mined disease information is incompatible with the disease information perceived by the user, it will eventually lead to the loss of users from the online medical consultation platform, degrading its operation and management. Using existing models to mine disease information leads to significant errors when users perceive the disease. Therefore, this research extends the latent Dirichlet allocation (LDA) and Twitter-LDA models to propose an intelligent topic model, PQDR-LDA. Compared with the Twitter-LDA model, the proposed model has a smaller perplexity value, stronger generalization ability, greater coherence value, lower correlation between topics, and stronger ability in extracting the disease information. It is found that the accuracy of disease diagnosis is very low, and the user's need for perceiving the disease will be reduced while using the traditional model to mine only the text of user questions on an online medical consultation platform. The accuracy of disease diagnosis does not decrease while only mining the doctor's reply text. Disease information that is more suitable for the consultation text can be obtained, which in fact cannot meet the user's real appeal for health, and reduces the users' needs in perceiving the disease. These findings have important management implications for the platform's operation and decision-making. Besides, users will ask questions in more medical texts simultaneously, which makes things more complicated. Unique management insights are obtained based on the disease information mining of user consultation texts through multiple consultation texts and multiple doctor replies.**

***Index Terms*—Big data analytics, data science in healthcare, healthcare technology, online medicine, PQDR-LDA model, text mining.**

Manuscript received 30 December 2021; revised 29 December 2022 and 29 May 2023; accepted 9 August 2023. The work of Zongrun Wang and Yanju Zhou was supported by the National Natural Science Foundation of China under Grant 72091515, Grant 71871232, and Grant 91846301. Review of this manuscript was arranged by Department Editor N. Gerdsri. *(Corresponding author: Zongrun Wang.)*

Xin Liu is with the College of Economics and Management, Hengyang Normal University of China, Hengyang 421007, China (e-mail: liunixs@tom.com).

Yanju Zhou and Zongrun Wang are with the School of Business, Central South University of China, Changsha 410017, China (e-mail: zhouyanju@csu.edu.cn; csu-wangzr@csu.edu.cn).

Ajay Kumar is with the EMLYON Business School, 69130 Ecully, France (e-mail: akumar@em-lyon.com).

Baidyanath Biswas is with the Trinity Business School, Dublin Trinity College, D02 F6N2 Dublin, Ireland (e-mail: baidyanath@gmail.com).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TEM.2023.3307550.

Digital Object Identifier 10.1109/TEM.2023.3307550

I. Introduction

ONLINE health care has been one of the most practical digital innovations since the creation of Web 2.0. Federico Sferrazz, the digital marketing manager of Daxue Consulting, commented that, "The development of online medical technology will alleviate the medical problems faced by the country" [1]. According to statistics from Rock Health, in 2021, a total of 808 digital health companies were invested in worldwide totaling an investment of $32 billion. A survey of user's inquiry texts were generated, collected and shared in a medical Q&A system of online medical business websites (e.g., "Ask a doctor quickly," "Seek medical advice," "Good doctor online," and "ask.39.net" in China, and "PatientsLikeMe," "DailyStrength," "Wellsphere," and "MDJunction" in foreign countries), and these websites have collected a large amount of data. Users with different health conditions and their family members participate in answering questions to describe their conditions, and medical personnel and experts are involved in answering question and providing opinions to users [2].

These data will ultimately exert significant impacts on the online healthcare operations. Given such an impressive prospect and the expected growth of services, more and more attention has been paid to online healthcare. Meanwhile, when observing the medical Q&A system, we found wide differences in medical concepts and perceptions of health information between medical professionals and users. Specifically, doctors master professional diagnosis and treatment and nursing knowledge, but the online medical platforms can only provide guiding ideas for the diagnosis and treatment information submitted by consultation users from the doctor's professional perspective, which is likely to deviate from the original online and offline diagnosis and treatment results, ideas, and needs of online consultation users. To figure out why, online consultation users make nothing of obscure medical concepts [3], [4], [5], leading to information distortion between users and doctors in understanding disease information, which is adverse to the growth of later-stage benefits. Therefore, it is very urgent to deeply understand and analyze online medical consultation platforms, and to extract the disease topics of users' online consultation from the interactive information of "inquiry" and "diagnosis" under these online medical platforms [6], [7]. In this article, we try to answer the question of how to most effectively acquire disease topic.

A. *Motivation*

This article is advanced through the inquiry environment of an online medical platform where users obtain medical information from inquiry texts to discern diseases. Like with many other online medical platforms, it is quite common for users to recognize the symptoms of a disease. For example, Li et al. [8] developed the medical knowledge extraction (MKE) system to extract medical crowd sourcing answers so that users can recognize diseases. The better that the disease recognized by users matches their symptoms, the more that they will rely on the online medical platform. Thus, this dependence will empower users to utilize more value-adding services through the platform to produce medical chain reactions, which can develop the incremental business of the platform. As a result, users receive lower prices, which is a win–win situation for users and the online platform. However, differentiating them from other online commercial platforms, the nature of disease recognition has brought unique challenges to online health care platforms. One of the challenges is mining for disease information. Receiving contrasting results from mining can lead to the distortion of users' recognition of diseases, thereby promoting the loss of potential users. For example, "online users will abandon an online medical platform if it cannot correctly recognize disease information" [9].

The mining of disease information plays a vital role in users' recognition of diseases. Various topic models can be used as data mining methods to analyze the potential texts. On a social platform, the analysis of the massive data generated by users and the acquisition of the topics of concern to users and their dynamic changes can provide assistance for public opinion analysis, personalized recommendations, accurate content delivery and other work [10], [11], [12]. However, using existing models to mine the disease information in users' online inquiry texts is often ineffective. Mining user questions alone sometimes cannot accurately obtain even the disease information expressed by the user. Mining doctors' answers alone sometimes cannot fully understand the health claims of users, thereby eventually reducing the demand for inquiry and damaging profitability of the platform because the medical terms in users' questions are sparse, and the text semantics are not clear; in contrast, the medical terms in doctors' answers are dense, and the text semantics are clear. Overall mining or key-point mining may be unable to obtain accurate information about a disease, thus forcing users and online medical platforms into a vicious circle. Therefore, it is necessary to expand the latent Dirichlet allocation (LDA) and Twitter-LDA topic models, that is, to realize short text mining, to combine users' questions and doctors' answers to improve the problem of text semantic fuzziness and entity sparsity, and to make the disease topics of online medical inquiry texts clearer to accurately express the users' health demands. In this article, the model relationship between mining disease information and users' questions and doctors' answers in an online medical inquiry environment is studied.

Another factor affecting the cognition of diseases is the increasing use of a single inquiry text and a single doctor's answer by platforms. Although a single question and answer may lead to more accurate results of disease cognition, if the results of mining disease information provides contrasting information, users will not recognize the disease in relation to their actual situation. This is often referred to as diagnosis risk in the literature [13]. To minimize the cost of this self-oriented high risk, users become increasingly dependent on raising questions through multiple inquiry texts and obtain answers from multiple doctors. For example, Adé et al. [14] believed that the future will be dominated by questions through multiple questions and answers. Mousavi et al. [15] suggested that this situation would rapidly intensify with an increase in the risk of mining disease information. Nearly 88% of users adopt the questioning strategy of multiple inquiry texts. Multiple questions and answers has brought more challenges and opportunities. From the user's perspective, a key strategy is to discern information from multiple doctors' answers by asking questions through multiple inquiry texts. Because the drugs described in each doctor's answer may be different, the treatment methods may be different, and the evaluation and testing may likewise differ. From the perspective of the medical platform, the challenge is to obtain the disease information more meticulously. Multiple doctors' answers to questions through multiple inquiry texts can put forward opinions and suggestions from different angles, which can describe the disease status together with the user's questions. In this article, the interaction between users who discern disease information and the medical platform mining disease information will be elaborated in the next section.

B. *Research Questions and Contributions*

Over the years, there has been controversy in regard to the number of doctors who provide answers to inquiry texts on online medical platforms. Users can usually send a series of questions through multiple inquiry texts each time they log in to an online medical platform, and multiple doctors can answer questions. Sometimes the high number of simultaneous questions cannot be handled by the online medical platform, and medical resources will be wasted if different doctors reply to the same questions of a user. Numerous questions and answers can lead to contrasting information about a disease through mining the online medical platform, thereby affecting users' understanding of diseases. Eventually, users may change medical platforms, especially when the cost of changing a medical platform is low. This raises an important management question: will users who submit questions through multiple inquiry texts and receive answers from multiple doctors develop conflicting understanding about a disease?

According to the above description, users who submit questions through multiple query texts and receive multiple doctors' answers should have contradictory understandings of a disease. Because multiple questions from users and answers from multiple doctors are both sparse and dense, the unstructured data are excessively rich, and the texts are too varied. This can easily lead to conflicting disease information. However, the results obtained through model optimization show that questions from multiple text inquiries and answers from multiple doctors may not be the reason for the contrasting information. It was found that under some conditions, mining disease information may still cause contradictions with questions from a single text and answers from a single doctor. This problem has an important impact for users who can discern information about diseases. For example, on the "Good Doctor" online medical platform, the process of a user asking questions from a single text and receiving answers from a single doctor is adopted to improve the effect of mining disease information and reduce the loss rate of users [16], [17]. However, our results demonstrate that this is not necessarily the optimal strategy. In this article, we try to elaborate on this important problem through model comparison.

Another significant attribute that affects users' cognition of diseases is the overall impact of user questions and doctor answers on mining disease information. With the reduction of user tolerance for contradictory disease information obtained through mining, a decline in the quality of mined information may reduce users' ability to discern disease information and may adversely affect users and online medical platforms. Dissatisfied users are more likely to abandon an online medical platform and seek better quality information from other online medical platforms. The research conducted by Lazard et al. [18] showed that users sensitive to the quality of disease cognition pay more attention to the effect of mining disease information than price or other factors. Users will eventually prefer online medical platforms with relatively high-quality mined data related to disease information. However, online medical platforms tend to mine disease information based on LDA and Twitter-LDA topic models [12]. Therefore, when users are more sensitive to the results of mining data related to disease information, can LDA and Twitter-LDA topic models provide better mining results?

The answer to the above question desired by the online medical platforms is "yes." By doing so, they can reduce technical costs to improve the effect of users' disease cognition by increasing the knowledge base capacity of online medical platforms. However, it is found that this is not entirely true. When the technical cost of online medical platforms decreases (mining with LDA and Twitter-LDA topic models), users will be adversely affected. In this article, we try to clarify this important problem through the conclusions obtained from the model comparisons, which is of important significance to management of online medical platforms in the adoption of model innovation as an advantage against market competition.

Therefore, the first problem to be solved in this article is: how to make the mined disease topic reflect the real health demands of users and reflect accurate disease information in the environment of multiconsultation texts and multidoctors' answers. The second problem to be solved is: how to mine clear and efficient disease topic in user's online medical consultation text where medical entities are both sparse and dense, and text semantics are both ambiguous and clear.

The remainder of this article is organized as follows. In section II, we described relevant work, and described the urgency and importance of this research through user data mining and topic mining of online medical texts in the online medical community. In section III, a disease theme model PQDR-LDA (Patients with questions and doctors respond Latent Dirichlet Allocation) based on the relationship between users' questions and doctors' answer is constructed. The model is then used for disease theme mining in users' online consultation text in the Q&A scenario of "120ask.com" website. In section IV, the experimental design is introduced in detail, and the preparation before experiment is described, including the analysis of experimental data set and the model evaluation indicators. Later, the experimental results are displayed and analyzed to obtain corresponding experimental conclusions. Finally, in Section V the full text is summarized and the direction of future work is proposed.

II. Related Work

A. Research on User Data Mining in Online Medical Communities

1) *Research on the Classification and Characteristics of Various Social Supports in Online Medical Communities:* Lucy et al. [19] studied negative moderation by the characteristics of shared information, including the amount of engagement, patients' precommitment and patients' social connectedness. Treatment effectiveness is closely related to community participants' perceptions about the treatment. Bar-Lev et al. [20] analyzed the content of the "emotional conversation" of HIV/AIDS support groups and demonstrated that emotional dynamics in online medical support groups are a moral concept rather than a simple psychological or therapeutic interactive component. Chuang et al. [21] used data from a 3-month time period on MedHelp (an online medical peer support community) to study the types of social support in the alcoholism community, especially the type of information support. Biyani et al. [22] showed that online medical users can gain experience from other online medical patients and use related medication or treatment advice as supporting information. Mo and Coulson [23] explored the social support nature of communication within online medical HIV/AIDS support groups.
2) *Factors Affecting the Interuser Exchange and Sharing of Health Information:* Heidelberger et al. [24] studied the main factors affecting the health decision-making behaviors of patients or doctors in online medical communities. Christensen et al. [25] examined the predictive factors of depression and anxiety on the MoodGYM website based on user characteristics, aiming to discover which factors predict changes in mental health status. Uden-Kraan et al. [26] participated in online medical support groups to understand the degree of suffering experienced by patients during treatment and to learn which processes occurred in these groups that were related to feelings of empowerment, thus proving the impact of participating in online medical support groups on empowerment. Selby et al. [27] compared the characteristics of smokers who did or did not publish data in online medical smoke cessation support groups, qualitatively analyzed the content of the discussion board, and determined the time needed by new users to accept feedback from existing members or forum managers.
3) *Research on the User Participation Modes in Medical Communities:* Tang and Yang [28] proposed a statistical

method that explores the quantitative relationship between messages published by users and replies from users to identify influential users in online medical communities. Mo et al. [29] explored the mechanism by which participating in an online support group (OSG) might facilitate 340 HIV/AIDS patients to obtain patient empowerment and found that the correlation between higher frequency of OSG use and more frequent empowerment processes can be measured based on the reception of useful information, the acceptance of social support, the discovery of positive meaning and assistance from others. Bao et al. [30] found that portal use is associated with improvements in patient health outcomes along multiple dimensions, including the frequency of hospital and emergency room visits, re-admission risk, and length of stay. Burri et al. [31] qualitatively analyzed all of the messages published by people who are trying to quit smoking in the online medical community in April 2005 and suggested that this online medical community was mainly used by women as an emotional support system and encouragement source in the weeks prior to trying to quit smoking.

In summary, users in online medical community communicate and share health information through different participation modes, and we can obtain users' thoughts and emotions from communication, and evaluate the classification and characteristics of social support. Even so, intelligent text processing methods have not been widely used in online medical community for a long time, and there are few studies on data mining of users' online consultation texts. The characteristics of the Internet without space–time limit enable users to express their subjective thoughts and emotions without scruples, making user's online consultation text an important medium for doctor–patient communication in the online medical community, as well as the most real reflection of users' information needs and the main place for doctors to answer. Hence, user's online consultation texts contain a large amount of user subjective information and medical-related information, which can reflect the subject characteristics that users pay attention to in the process of medical treatment, and more demand systematic and scientific methods to conduct in-depth theme mining on user online consultation texts.

B. Research on Topic Mining in Online Medical Texts

With the promotion and rapid development of "Internet+," users have been gradually changing their medical advice-seeking patterns [32]. Based on the increasing popularity of medical information platforms and online health communities as well as the boom in online medical texts, many researchers have found that these texts can be recognized as important data sources to identify medical service needs from the user perspective [15], [33], [34]. Emmert et al. [35] conducted a quantitative analysis on comments from 3000 texts on Jameda (a German healthcare platform) and found that the topics of user comments were summarized by doctors, office staff and other related factors (such as waiting time and equipment). Jung et al. [36] used text mining technology to identify the key topics of hospital services on two health community platforms (Naver and Daum) in South Korea, including service, environment, professionalism, impression, process, and popularity. Ranard et al. [37] compared the text reviews on Yelp (an online review platform) with the HCAHPS (the Hospital Consumer Assessment of Healthcare Providers and Systems), an American standard for the evaluation of patient visits, and found that the focus topics in the former included most of the aspects of the HCAHPS; these research results can help decision-makers determine what users value most about hospital quality. Bekhuis et al. [38] used NLP-related technologies to extract disease-related topic words from online medical platform postings and used them as a basis for effective organization and classification. Chen et al. [39] proposed a clustering method for postings of online medical community platforms (such as diabetes, fibromyalgia, and breast cancer platforms) and found that popular topics are distributed differently on different platforms.

Lu et al. [40] incorporated the specific characteristics of the medical field into the text clustering method to explore health-related topics in the online medical community and used these topics to understand patients' interests and needs. Attard and Coulson [41] analyzed the topic information exchanged in the Parkinson's disease forum to gain insight into the positive and negative aspects of online communication. Portier et al. [42] applied text sentiment analysis and topic models to CSN breast and colorectal cancer discussion posts (2005 to 2010) to explore the emotional changes of the post initiators and used the topics discussed to divide social support standards. Chen et al. [39] used a semiautomated method to identify the differences in the topic content (such as the type of support, medications, and how to manage self-health, etc.) in various situations on the online medical forum.

Zeng et al. [43] introduced the unified medical language system (UMLS) to extract relevant features in the medical field and used machine learning methods to analyze disease-related topics on diabetes community platforms, effectively mining potential information related to diabetes. Hao et al. [44] used the LDA topic model to mine and analyze 100 000 texts from "haoddf.com" (an online medical platform), compared with reviews about pediatric and OB departments on RateMD.com (an American website), and found that nondoctoral employees (such as nurses) in American medical institutions play an important role in online medical services. Lu et al. [40] used UMLS to extract medical features, clustered the popular topics of online medical community platforms, and analyzed the popular topics by machine learning. Zhai et al. [45] used text mining methods on the diagnosis and treatment-related online texts on online traditional Chinese medicine community platforms, built a biological network of two symptoms of Qi deficiency (deficiency of vital energy) and Qi stagnation (stagnation of the circulation of vital energy), and mined available information related to the symptoms. Ruan et al. [46] studied the eight most popular online medical community platforms in China and integrated a Chinese medical knowledge base that could be used for Chinese text study. Fang et al. [47] conducted research on the domestic online medical community platform DXY.com and used an unsupervised transfer learning method to automatically assign users with different needs to corresponding medical experts, which is conducive to efficient communication in the community. Yang et al. [48] used the LDA model to automatically identify hidden topics and words related to every topic in the information published by users in the private and public sections of the online cancer community.

In summary, topic mining of online medical texts has achieved remarkable results, but there are currently few studies on the topic mining of disease information in online medical users' inquiry texts. More importantly, user questions and doctor answers contain considerable medical information and subjective user information. However, the traditional LDA model is suitable for long texts using standard language and assumes that the texts are composed of a set of disordered words, ignoring the contextual connection. Despite various improved LDA topic models, there is no model for the mining of disease information of users' texts on online medical community platforms. In addition, the improved models do not consider the relevant word distribution features and text architectural features in texts, and they are not suitable for the systematic analysis of data topics in the users' questions and doctors' answers. Therefore, it is necessary to extend the original topic model and make it better adaptable to the features of texts.

In this article, the existing LDA and Twitter-LDA models were expanded to propose a disease topic model PQDR-LDA based on the relationship between users' questions and doctors' answers. The model was used to mine topics related to diseases, clarify the text and present users' health appeals in online medical inquiries, which are characterized by sparse and dense medical entities and fuzzy and clear text semantics. First, the probability map of the model and the document generation process were sorted. Specifically, when analyzing user questions, the processing method of the Twitter-LDA model was retained. A user question contains only one topic from the distribution of users' topics, and there are background topics to filter the influence of high-frequency background words. When analyzing a doctor's answer, which contains only one topic and is consistent with the topic of the user's question, there is additional information to supplement the user's question about a disease. Second, on this basis, the Gibbs sampling method was used to derive the model and obtain the estimation methods for the model parameters and the Gibbs sampling algorithm. Finally, to verify the validity of the PQDR-LDA model, the model was applied to the medical inquiry text dataset of the online medical publishers of the respiratory medicine department in the Q&A scenario of "120ask.com" and was compared with the Twitter-LDA model in the perplexity indicator and the coherence indicator, which is likewise suitable for disease theme mining scenarios on the online medical community platform.

III. Disease Topic Mining of Users' Inquiry Texts With Online Medical Data

A. Model Description and Symbol Explanation

Users' online medical inquiry text data from online medical publishers are characterized by a major proportion of paragraph-based, narrative languages, a large amount of data, a sparsity of medical entities, an abundance of unstructured data, and the mixture and complexity of the texts. The content of the texts fail to sufficiently reflect the online medical entities and content desired by the users. Therefore, the medical entities identified in the texts are summarized and used as the users' online medical inquiry texts. However, in the network platform framework of texts, users' questions and doctors' answers (Q&As) are full of various diseases, diagnosis and treatment information, and existing topic models usually lead to unsatisfactory mining results. Therefore, it is necessary to find a way to mine disease information from questions, and medical condition analysis and recommendations from answers to make user questions clearer. Based on the features of the users' questions and doctors' answers in texts, a new topic model, the PDQR-LDA model (patients with questions and doctors respond latent Dirichlet allocation), is proposed based on the LDA and Twitter-LDA models and is used to mine vague disease topics in the users' online medical inquiry texts.

Under the PQDR-LDA model, every user's text data is divided into two parts: the user's question (Q) and the doctor's answer (A). Therefore, the PQDR-LDA model is also divided into two parts for the processing texts: Q processing and A processing. In Q processing, the short text processing method of the Twitter-LDA model is retained; namely, it is assumed that there is only one subject disease in every user's question, and that all the words in the user's question share this subject, which is called the disease topic. The overall text disease topics come from the distribution of the users' disease topics. Meanwhile, the Q part of users' texts is set with a background topic to filter some high-frequency background words that are not related to the disease topics. The A part of the texts shares its disease topic with the corresponding Q, and the words used by the doctors' answers and users' questions come from the distribution of the same disease topic words. Since there are usually multiple answers corresponding to one question, this portion needs to select the users' questions and the accepted doctors' answers. Since there may be supplementary disease information in the doctors' answers, additional topics are introduced in the model to obtain the supplementary words of accepted doctors' answers. In addition, this model assumes 1) in the data from online medical publishers, the doctors' answers accepted for different users' texts are not the same, so additional topics are set for them; therefore, they are different in word distribution; and 2) users with the same age, gender, and city in the medical inquiry text data of the online medical publisher are the same user.

After learning, the model can mine the distribution of the users' disease topics as well as the disease topic of every text and learn the additional topics of the accepted answers from doctors. The mined disease topics of texts can be used as the basis for obtaining the users' disease topics, and the additional topics of the accepted answers from doctors can be taken as a supplementary basis for the users' disease topics to obtain accurate users' disease topics for subsequent analysis. There are many mathematical symbols in the probability graph and model derivation process of the subsequent PQDR-LDA model. Therefore, it is necessary to introduce and explain the mathematical symbols. The explanation of the mathematical symbols is briefly given in Appendix A.

B. Construction of a Disease Topic Mining Model for User Online Medical Inquiry Texts

1) Model Framework: In the PQDR-LDA model, the topic of each user's online medical inquiry texts can be expressed as a distribution of multiple disease topics $\overrightarrow{\theta_u}$, $\overrightarrow{\theta_u} \sim \mathrm{Dir}(\overrightarrow{\alpha})$, and the component of $\overrightarrow{\theta_u}$ represents the probability of multiple disease topics being present in a user's texts. The topics of users' questions, which are a part of an online medical inquiry text, come from the distribution of the users' disease topics, and each text question contains only one topic. In addition, the disease topic t of every text can be expressed as the word distribution $\overrightarrow{\phi_t}$, $t \in [1, 2, \ldots, T]$ of V words. Since the users' questions include background topics, every word in a user's question belongs to either the disease topic or the user's background topic. The word distribution of the background topics is $\overrightarrow{\phi_b}$, where $\overrightarrow{\phi_b} \sim \mathrm{Dir}(\overrightarrow{\beta_b})$; the component of $\overrightarrow{\phi_b}$ represents the probability of a word appearing in the background topics of a text. The disease topic of texts is $\overrightarrow{\phi_t}$, where $\overrightarrow{\phi_t} \sim \mathrm{Dir}(\overrightarrow{\beta})$; the component of $\overrightarrow{\phi_t}$ represents the probability of the word appearing in the disease topics t of a text. Whether a word in a user's question belongs to the disease topics or the background topics depends on the correlation distribution of the background topics $\overrightarrow{\pi_b}$, $\overrightarrow{\pi_b} \sim \mathrm{Dir}(\overrightarrow{\lambda_b})$; the component π_b^0 of $\overrightarrow{\pi_b}$ represents the probability of the background topics, and π_b^1 represents the probability of the disease topics in a text.

The topics of the doctors' answers, which are part of the online medical inquiry text, come from the distribution of the users' topics, as the users' questions and the doctors' answers share the same disease topics. However, there will typically be additional topics in the doctors' answers, so the topic of every word in a doctor's answer belongs to either the disease topic or the users' additional topics. The word distribution of the additional topics is $\overrightarrow{\phi_{ex,u}}$, where $\overrightarrow{\phi_{ex,u}} \sim Dir(\overrightarrow{\beta_{ex}})$; the component of $\overrightarrow{\phi_{ex,u}}$ represents the probability of the word appearing in the additional topics under users u. Whether a word in a doctor's answer belongs to the disease topic or additional topics depends on the correlation distribution of the additional topics $\overrightarrow{\pi_{ex}}$, $\overrightarrow{\pi_{ex}} \sim Dir(\overrightarrow{\lambda_{ex}})$; the component π_{ex}^0 of $\overrightarrow{\pi_{ex}}$ represents the probability of the additional topics, and π_{ex}^1 represents the probability of the disease topic in the texts.

In summary, the probability graph of the PQDR-LDA model is shown in Fig. 1.

In the graph, the white circles represent the hidden variables; the gray circles represent the observable variables; the boxes represent cyclic relationships; the letters in the lower right corner of the boxes represent the numbers of cycles; and the arrows represent the conditional dependence relationship between the variables. The meanings of the symbols in the graph are given in Table I.

2) Model Derivation: After the PQDR-LDA model is established, the medical inquiry text dataset of online medical publishers can be used for model learning and to obtain the

TABLE I
Online Medical Inquiry Text Generation Process Based on the PQDR-LDA Model

PQDR-LDA model: PQDR-LDA models the association relationship between users' questions and doctors' answers, and introduces additional topics to determine disease-related information.

1. The prior parameter $\overrightarrow{\beta_b}$ follows the Dirichlet distribution, and generates the word distribution of the background topics $\overrightarrow{\phi_b} \sim Dir\left(\overrightarrow{\beta_b}\right)$; $\overrightarrow{\phi_b}$ follows the multinomial distribution;

2. The prior parameter $\overrightarrow{\lambda_b}$ follows the Dirichlet distribution, and generates the correlation distribution of the background topics $\overrightarrow{\pi_b} \sim Dir\left(\overrightarrow{\lambda_b}\right)$; $\overrightarrow{\pi_b}$ follows the multinomial distribution;

3. The prior parameter $\overrightarrow{\lambda_{ex}}$ follows the Dirichlet distribution, and generates the correlation distribution of the additional topics $\overrightarrow{\pi_{ex}} \sim Dir\left(\overrightarrow{\lambda_{ex}}\right)$; $\overrightarrow{\pi_{ex}}$ follows the multinomial distribution;

4. For the disease topic t of every text, $t \in \left[1, 2, \cdots, T\right]$:

The prior parameter $\overrightarrow{\beta}$ follows the Dirichlet distribution, and generates the word distribution of the disease topics of the texts $\overrightarrow{\phi_t} \sim Dir\left(\overrightarrow{\beta}\right)$; $\overrightarrow{\phi_t}$ follows the multinomial distribution;

5. For every user u in the medical publishers' dataset, $u \in \left[1, U\right]$:

(1) The prior parameter $\overrightarrow{\alpha}$ follows the Dirichlet distribution, and generates the user's disease topic distribution $\overrightarrow{\theta_u} \sim Dir\left(\overrightarrow{\alpha}\right)$; $\overrightarrow{\theta_u}$ follows the multinomial distribution;

(2) The prior parameter $\overrightarrow{\beta_{ex}}$ follows the Dirichlet distribution, and generates the word distribution of user additional topics $\overrightarrow{\phi_{ex,u}} \sim Dir\left(\overrightarrow{\beta_{ex}}\right)$; $\overrightarrow{\phi_{ex,u}}$ follows the multinomial distribution;

(3) For every text m of the user u, $m \in \left[1, D_u\right]$:

(a) The disease topic z_m of text m is selected from $\overrightarrow{\theta_u}$ (the disease topic distribution of the user u), $z_m \sim Multi\left(\overrightarrow{\theta_u}\right)$;

(b) For every word n in the users' question in a text, $n \in \left[1, N_m^b\right]$;

i. The word correlation $Y_{u,m,n}^b$ of the background topics is selected from the background topic correlation distribution $\overrightarrow{\pi_b}$, $Y_{u,m,n}^b \sim Multi\left(\overrightarrow{\pi_b}\right)$;

ii. If the word correlation $Y_{u,m,n}^b = 0$, then the word is related to the background topics, and it is needed to select the word $w_{u,m,n}^b$ from the background topic word distribution $\overrightarrow{\phi_b}$, $w_{u,m,n}^b \sim Multi\left(\overrightarrow{\phi_b}\right)$; if the word correlation $Y_{u,m,n}^b = 1$, then the word is not related to the background topics, and it is needed to select the word $w_{u,m,n}^b$ from the word distribution $\overrightarrow{\phi_{z_m}}$ of disease topics z_m, $w_{u,m,n}^b \sim Multi\left(\overrightarrow{\phi_{z_m}}\right)$.

(c) For every word n in the doctors' answers in a text m, $n \in \left[1, N_m^{ex}\right]$:

i. The word correlation of the additional topics $Y_{u,m,n}^{ex}$ is selected from the additional topics correlation distribution $\overrightarrow{\pi_{ex}}$, $Y_{u,m,n}^{ex} \sim Multi\left(\overrightarrow{\pi_{ex}}\right)$;

ii. If the word correlation $Y_{u,m,n}^{ex} = 0$, then the word is related to the additional topics, and it is needed to select the word $w_{u,m,n}^{ex}$ from the additional topics word distribution $\overrightarrow{\phi_{ex,u}}$; if the word correlation $Y_{u,m,n}^{ex} = 1$, then the word is not related to the additional topic, and it is needed to select the word $w_{u,m,n}^{ex}$ from the word distribution $\overrightarrow{\phi_{z_m}}$ of the online medical inquiry texts disease topics z_m, $w_{u,m,n}^{ex} \sim Multi\left(\overrightarrow{\phi_{z_m}}\right)$.

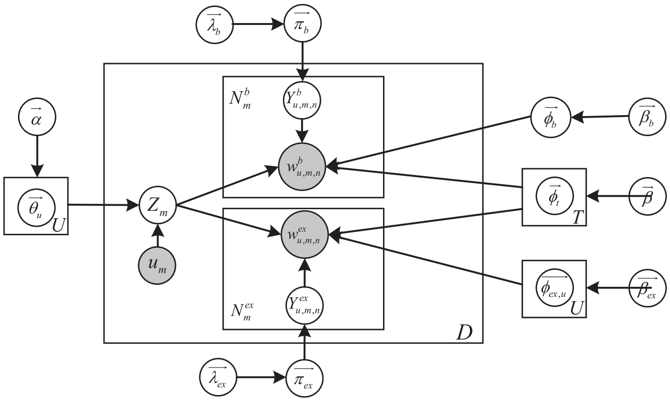

$\overrightarrow{\lambda_b}$
$\overrightarrow{\pi_b}$
$\vec{\alpha}$
N_m^b
$Y_{u,m,n}^b$
$w_{u,m,n}^b$
$\overrightarrow{\phi_b}$
$\overrightarrow{\beta_b}$
$\overrightarrow{\theta_u}$
U
Z_m
$\overrightarrow{\phi_t}$
T
$\vec{\beta}$
$w_{u,m,n}^{ex}$
u_m
$\overrightarrow{\phi_{ex,u}}$
U
$\overrightarrow{\beta_{ex}}$
N_m^{ex}
$Y_{u,m,n}^{ex}$
D
$\overrightarrow{\lambda_{ex}}$
$\overrightarrow{\pi_{ex}}$

Fig. 1. Model probability graph of PQDR-LDA.

parameters in the model. The PQDR-LDA model is based on Bayesian theory, and all of the parameters of the model are random variables with prior distributions. In the PQDR-LDA model, the main parameters to be obtained include $\overrightarrow{\theta}_u$, $\overrightarrow{\phi_t}$, $\overrightarrow{\phi_b}$, $\overrightarrow{\phi_{ex}}$, $\overrightarrow{\pi_b}$, and $\overrightarrow{\pi_{ex}}$. Solving the model identifies the posterior distribution through the prior distributions of these parameters and the sample information of the dataset. However, due to the existence of the hidden variables (the disease and word topics that cannot be known), it is difficult to accurately calculate the parameters. For this kind of calculation problem, there are two common solutions: variational inference and sampling estimation. Compared with variational inference, sampling estimation is more commonly used due to its simple and direct processing process. The derivation of the PQDR-LDA model in this article is based on the Gibbs sampling method. The probability distribution model is briefly given in Appendix C.

a) Probability distribution of disease topics in online medical inquiry texts: As seen from the model definition, the distribution of user disease topics $\overrightarrow{\theta_u}$ in online medical publishers' medical inquiry text dataset follows the multinomial distribution, and the prior parameter of $\overrightarrow{\theta_u}$ is $\overrightarrow{\alpha}$, which follows the Dirichlet distribution, so that the posterior distribution of $\overrightarrow{\theta_u}$ can be obtained from the sample information. Because the disease topics of user $u$come from $\overrightarrow{\theta_u}$ and because $\overrightarrow{\theta_u}$ corresponds to the inquiry texts (namely, every text has only one disease topic), the sample information obtained by $\overrightarrow{\theta_u}$ under the multinomial distribution can be obtained by counting the number of texts.

It is assumed that $\overrightarrow{z_u} = (z_u^1, z_u^2, \cdots, z_u^{D_u})$, which is used to represent the disease topics of all online medical inquiry texts from user u, and the generation probability of $\overrightarrow{z_u}$ is

$$p\left(\overrightarrow{z_u} \,|\, \overrightarrow{\alpha}\,\right) = \frac{\Delta\left(\overrightarrow{\alpha} + \overrightarrow{n_u}\right)}{\Delta\left(\overrightarrow{\alpha}\right)} \tag{1}$$

where $\overrightarrow{n_u} = \{n_u^t\}_{t=1}^T$ and n_u^t represents the number of occurrences of disease topic t in all texts from user u. Additionally, since the users are independent of each other, the generation probability of the disease topics is as follows:

$$p\left(\overrightarrow{z} \,|\, \overrightarrow{\alpha}\,\right) = \prod_{u=1}^{U} \frac{\Delta\left(\overrightarrow{\alpha} + \overrightarrow{n_u}\right)}{\Delta\left(\overrightarrow{\alpha}\right)} \tag{2}$$

b) Topic probability distribution of the users' questions: Because user questions include background topics and disease topics, when generating every word in the Q part of the data, it is necessary to first determine whether a word belongs to the background topics or the disease topic. If it is a background topic, a word is selected from the word distribution; if not, a word is selected from the word distribution of the disease topics. Whether a word is a background or disease topic is determined by the multinomial distribution $\overrightarrow{\pi_b}$, and the prior parameter of $\overrightarrow{\pi_b}$ is $\overrightarrow{\lambda_b}$, which follows the Dirichlet distribution. Therefore, the posterior distribution of $\overrightarrow{\pi_b}$ can be obtained through sample information. Since the Q parts all share the same distribution $\overrightarrow{\pi_b}$ and $\overrightarrow{\pi_b}$ corresponds to words, the sample information needs to be detailed to words, and the words should be counted.

It Is assumed that $\overrightarrow{Y_b} = (Y_b^1, Y_b^2, \cdots, Y_b^{N_b})$ is used to determine whether all of the words in the Q part of the data are related to the background topics, and N_b represents the total number of words in the Q part. Then, the generation probability of $\overrightarrow{Y_b}$ is as follows:

$$p\left(\overrightarrow{Y_b} \,\middle|\, \overrightarrow{\lambda_b}\right) = \frac{\Delta\left(\overrightarrow{\lambda_b} + \overrightarrow{R_b}\right)}{\Delta\left(\overrightarrow{\lambda_b}\right)} \tag{3}$$

where $\overrightarrow{R_b} = \{R_b^r\}_{r=1}^2$, and $R_b^1 + R_b^2 = N_b$; R_b^1 represents the number of the occurrence of the background topics in the Q part of the data (the number of words belonging to the background topic distribution), and R_b^2 represents the number of the occurrence of disease topics in the Q part of the data (the number of words belonging to the disease topic distribution).

Using Bernoulli's notation, the probability distributions of disease topics and background topics are considered as follows:

$$p\left(\overrightarrow{Z_b} \,\middle|\, \overrightarrow{\alpha}, \overrightarrow{\lambda_b}\right) = \left[\frac{\Delta\left(\overrightarrow{\lambda_b} + \overrightarrow{R_b^B}\right)}{\Delta\left(\overrightarrow{\lambda_b}\right)}\right]^{1-e}$$
$$\times \left[\frac{\Delta\left(\overrightarrow{\lambda_b} + \overrightarrow{R_b^{\neg B}}\right)}{\Delta\left(\overrightarrow{\lambda_b}\right)} \times \prod_{u=1}^{U} \frac{\Delta\left(\overrightarrow{\alpha} + \overrightarrow{n_u}\right)}{\Delta\left(\overrightarrow{\alpha}\right)}\right]^{e}. \tag{4}$$

In the formula, the background topics are marked with B, and the disease topics are marked with $\neg B$. When $e = 0$, the formula expresses the probability distribution of background topics; when $e = 1$, the formula expresses the probability distribution of disease topics in the Q part of the data.

c) Topic probability distribution of the doctors' answers: Since the doctor answers contain both additional topics and disease topics, when generating every word in the A part of the data, it is necessary to first determine whether the word belongs to the additional topics or disease topics. If it is an additional topic, a word is selected from the word distribution; if not, a word is selected from the word distribution of the disease topics. Whether a word is an additional topic is determined by the multinomial distribution $\overrightarrow{\pi_{ex}}$, and the prior parameter of $\overrightarrow{\pi_{ex}}$ is $\overrightarrow{\lambda_{ex}}$, which follows the Dirichlet distribution. Therefore, the posterior distribution of $\overrightarrow{\pi_{ex}}$ can be obtained through sample information. Since the texts of the A part of the data share the same distribution $\overrightarrow{\pi_{ex}}$ and $\overrightarrow{\pi_{ex}}$ correspond to words, the sample information needs to be detailed to words, and the words should be counted.

It is assumed that $\overrightarrow{Y_{ex}} = (Y_{ex}^1, Y_{ex}^2, \cdots, Y_{ex}^{N_{ex}})$ is used to determine whether all the words of the A part of the data are related to the additional topics, and N_{ex} represents the total number of words in the A part of the data. Then, the generation probability of $\overrightarrow{Y_{ex}}$ is as follows:

$$p\left(\overrightarrow{Y_{ex}}\middle|\overrightarrow{\lambda_{ex}}\right) = \frac{\Delta\left(\overrightarrow{\lambda_{ex}} + \overrightarrow{R_{ex}}\right)}{\Delta\left(\overrightarrow{\lambda_{ex}}\right)} \tag{5}$$

where $\overrightarrow{R_{ex}} = \{R_{ex}^r\}_{r=1}^2$, and $R_{ex}^1 + R_{ex}^2 = N_{ex}$. R_{ex}^1 represents the number of occurrences of the additional topics in the A part of the data (the number of words belonging to the additional topics distribution), and R_{ex}^2 represents the number of occurrences of the disease topics in the A part of the data (the number of words belonging to disease topic distribution).

Using Bernoulli's notation, the probability distributions of disease topics and additional topics are considered as follows:

$$\begin{aligned} &p\left(\overrightarrow{Z_{ex}}\middle|\overrightarrow{\alpha}, \overrightarrow{\lambda_{ex}}\right) \\ &= \left(\frac{\Delta\left(\overrightarrow{\lambda_{ex}} + \overrightarrow{R_{ex}^{E}}\right)}{\Delta\left(\overrightarrow{\lambda_{ex}}\right)}\right)^{1-ex} \\ &\quad \times \left[\frac{\Delta\left(\overrightarrow{\lambda_{ex}} + \overrightarrow{R_{ex}^{\neg E}}\right)}{\Delta\left(\overrightarrow{\lambda_{ex}}\right)} \times \prod_{u=1}^{U} \frac{\Delta\left(\overrightarrow{\alpha} + \overrightarrow{n_u}\right)}{\Delta\left(\overrightarrow{\alpha}\right)}\right]^{ex}. \end{aligned} \tag{6}$$

In the formula, additional topics are denoted with E, and the disease topics are marked as $\neg E$. When $ex = 0$, the formula expresses the probability distribution of the additional topics; when $ex = 1$, the formula expresses the probability distribution of the disease topics in the A part of the data.

d) Probability distribution of words: After obtaining the disease topics $\overrightarrow{z}$, the background topics of the users' questions $\overrightarrow{Y_b}$, and the additional topics of the doctors' answers $\overrightarrow{Y_{ex}}$, the words in the data can be generated according to $\overrightarrow{z}$, $\overrightarrow{Y_b}$ and $\overrightarrow{Y_{ex}}$. Therefore, all the words in the data are divided into disease topics, additional topics, and background topics.

Users' disease topics t, are found by obtaining $\overrightarrow{\phi_t}$ (which follows the multinomial distribution) from the prior parameter $\overrightarrow{\beta}$ (which follows the Dirichlet distribution), and the words are generated and selected from $\overrightarrow{\phi_t}$. Therefore, the sample information can be used to acquire the posterior distribution of $\overrightarrow{\phi_t}$. However, texts of the same disease topics share the same word distribution $\overrightarrow{\phi_t}$, and $\overrightarrow{\phi_t}$ corresponds to words. The sample information needs to be detailed to words, and the words should be counted.

It is assumed that the topics t selected from the data are consistent with the users' disease topics, the word vector $\overrightarrow{W_t}$ is formed, and the generation probability of $\overrightarrow{W_t}$ is

$$p\left(\overrightarrow{W_t}\middle|\overrightarrow{\beta}\right) = \frac{\Delta\left(\overrightarrow{\beta} + \overrightarrow{n_t}\right)}{\Delta\left(\overrightarrow{\beta}\right)} \tag{7}$$

where $\overrightarrow{n_t} = \{n_t^v\}_{v=1}^V$, and n_t^v represents the number of occurrences of word v in the disease topics t from the data. Because the disease topics of the users' texts are independent of each other, the generation probability of all words under the disease topic is

$$p\left(\overrightarrow{W_D}\middle|\overrightarrow{\beta}\right) = \prod_{t=1}^{T} \frac{\Delta\left(\overrightarrow{\beta} + \overrightarrow{n_t}\right)}{\Delta\left(\overrightarrow{\beta}\right)} \tag{8}$$

where $\overrightarrow{W_D}$ represents the word vector composed of all the words from the disease topics in the data.

Using the same method as above to process background topics can also generate the generation probability for all the background topic words in the data

$$p\left(\overrightarrow{W_b}\middle|\overrightarrow{\beta_b}\right) = \frac{\Delta\left(\overrightarrow{\beta_b} + \overrightarrow{n_b}\right)}{\Delta\left(\overrightarrow{\beta_b}\right)} \tag{9}$$

where $\overrightarrow{n_b} = \{n_b^v\}_{v=1}^V$, n_b^v represents how many times word v under the background topics appears in the data.

When using the same method as above to process background topics, because the additional topics are distributed differently in users and the users are independent of each other, the generation probability for all the additional topic words from the data is formed as

$$p\left(\overrightarrow{W_{ex}}\middle|\overrightarrow{\beta_{ex}}\right) = \prod_{u=1}^{U} \frac{\Delta\left(\overrightarrow{\beta_{ex}} + \overrightarrow{n_{ex,u}}\right)}{\Delta\left(\overrightarrow{\beta_{ex}}\right)} \tag{10}$$

where $\overrightarrow{n_{ex,u}} = \{n_{ex,u}^v\}_{v=1}^V$, and $n_{ex,u}^v$ is how many times word v under the additional topics of user u appears in the data.

Since the topics are independent of each other, the generation probability of all the words in the data is

$$\begin{aligned} &p\left(\overrightarrow{W}\middle|\overrightarrow{Z}, \overrightarrow{Y_b}, \overrightarrow{Y_{ex}}, \overrightarrow{\beta}, \overrightarrow{\beta_b}, \overrightarrow{\beta_{ex}}\right) \\ &= \left[\prod_{t=1}^{T} \frac{\Delta\left(\overrightarrow{\beta} + \overrightarrow{n_t}\right)}{\Delta\left(\overrightarrow{\beta}\right)}\right] \times \frac{\Delta\left(\overrightarrow{\beta_b} + \overrightarrow{n_b}\right)}{\Delta\left(\overrightarrow{\beta_b}\right)} \\ &\quad \times \left[\prod_{u=1}^{U} \frac{\Delta\left(\overrightarrow{\beta_{ex}} + \overrightarrow{n_{ex,u}}\right)}{\Delta\left(\overrightarrow{\beta_{ex}}\right)}\right]. \end{aligned} \tag{11}$$

e) Joint probability distribution of topics and words:

$$\begin{aligned} &p\left[\overrightarrow{W}, \overrightarrow{Z_b}, \overrightarrow{Z_{ex}}\middle|\overrightarrow{\alpha}, \overrightarrow{\lambda_b}, \overrightarrow{\lambda_{ex}}, \overrightarrow{\beta}, \overrightarrow{\beta_b}, \overrightarrow{\beta_{ex}}\right] \\ &= \left[\frac{\Delta\left(\overrightarrow{\lambda_b} + \overrightarrow{R_b}\right)}{\Delta\left(\overrightarrow{\lambda_b}\right)} \times \frac{\Delta\left(\overrightarrow{\lambda_{ex}} + \overrightarrow{R_{ex}}\right)}{\Delta\left(\overrightarrow{\lambda_{ex}}\right)} \times \prod_{u=1}^{U} \frac{\Delta\left(\overrightarrow{\alpha} + \overrightarrow{n_u}\right)}{\Delta\left(\overrightarrow{\alpha}\right)}\right] \end{aligned}$$

TABLE II
ESTIMATED VALUES OF THE PARAMETERS

$\theta_u^t = \frac{\alpha^t + n_u^t}{\sum_{t=1}^{T}\left(\alpha^t + n_u^t\right)}$	$\phi_t^v = \frac{\beta^v + n_t^v}{\sum_{v=1}^{V}\left(\beta^v + n_t^v\right)}$
$\pi_b^r = \frac{\lambda_b^r + R_b^r}{\sum_{r=1}^{2}\left(\lambda_b^r + R_b^r\right)}$	$\pi_{ex}^r = \frac{\lambda_{ex}^r + R_{ex}^r}{\sum_{r=1}^{2}\left(\lambda_{ex}^r + R_{ex}^r\right)}$
$\phi_b^v = \frac{\beta_b^v + n_b^v}{\sum_{v=1}^{V}\left(\beta_b^v + n_b^v\right)}$	$\phi_{ex,u}^v = \frac{\beta_{ex}^v + n_{ex,u}^v}{\sum_{v=1}^{V}\left(\beta_{ex}^v + n_{ex,u}^v\right)}$

$$\times\left[\prod_{t=1}^{T}\frac{\Delta\left(\overrightarrow{\beta}+\overrightarrow{n_t}\right)}{\Delta\left(\overrightarrow{\beta}\right)}\times\frac{\Delta\left(\overrightarrow{\beta_b}+\overrightarrow{n_b}\right)}{\Delta\left(\overrightarrow{\beta_b}\right)}\times\prod_{u=1}^{U}\frac{\Delta\left(\overrightarrow{\beta_{ex}}+\overrightarrow{n_{ex,u}}\right)}{\Delta\left(\overrightarrow{\beta_{ex}}\right)}\right]. \tag{12}$$

3) Parameter Estimation: The Gibbs sampling conditional probability acquisition method in Appendix D is used to sample topics from the texts or words and to obtain $\overrightarrow{Z}$. When it reaches a stationary distribution, it is the desired sample set. Then, we can use the conjugate principle of the Dirichlet distribution and the multinomial distribution to obtain the posterior probability with the parameters as follows:

$$\mathrm{Dir}\left(\overrightarrow{p}\,|\overrightarrow{\alpha}\,\right)+\mathrm{MultiCount}\left(\overrightarrow{m}\right)=\mathrm{Dir}\left(\overrightarrow{p}\,|\overrightarrow{\alpha}+\overrightarrow{m}\,\right). \tag{13}$$

Therefore, the posterior distribution of the PQDR-LDA model with various parameters can be obtained as follows:

$$\begin{aligned}
&p(\overrightarrow{\theta_u}|\overrightarrow{n_u},\overrightarrow{\alpha})=\mathrm{Dir}(\overrightarrow{\theta_u}|\overrightarrow{\alpha}+\overrightarrow{n_u}) && p(\overrightarrow{\pi_b}|\overrightarrow{R_b},\overrightarrow{\lambda_b})=\\
& && \mathrm{Dir}(\overrightarrow{\pi_b}|\overrightarrow{\lambda_b}+\overrightarrow{R_b})\\
&p(\overrightarrow{\pi_{ex}}|\overrightarrow{R_{ex}},\overrightarrow{\lambda_{ex}}) && p(\overrightarrow{\phi_t}|\overrightarrow{n_t},\overrightarrow{\beta})\\
&=\mathrm{Dir}(\overrightarrow{\pi_{ex}}|\overrightarrow{\lambda_{ex}}+\overrightarrow{R_{ex}}) && =\mathrm{Dir}(\overrightarrow{\phi_t}|\overrightarrow{\beta}+\overrightarrow{n_t})\\
&p(\overrightarrow{\phi_b}|\overrightarrow{n_b},\overrightarrow{\beta_b}) && p(\overrightarrow{\phi_{ex,u}}|\overrightarrow{n_{ex,u}},\overrightarrow{\beta_{ex}})\\
&=\mathrm{Dir}(\overrightarrow{\phi_b}|\overrightarrow{\beta_b}+\overrightarrow{n_b}) && =\mathrm{Dir}(\overrightarrow{\phi_{ex,u}}|\overrightarrow{\beta_{ex}}+\overrightarrow{n_{ex,u}}).
\end{aligned}$$

The parameters are estimated according to the posterior distribution with the parameters. When the posterior distribution is $\mathrm{Dir}(\overrightarrow{p}|\vec{\alpha}+\vec{n})$, the expected value of $\overrightarrow{p}$ is as the following example:

$$E\left(\overrightarrow{p}\right)=\left[\frac{\alpha^1+n_u^1}{\sum_{t=1}^{T}\left(\alpha^t+n_u^t\right)},\frac{\alpha^2+n_u^2}{\sum_{t=1}^{T}\left(\alpha^t+n_u^t\right)},\cdots\frac{\alpha^t+n_u^t}{\sum_{t=1}^{T}\left(\alpha^t+n_u^t\right)}\right]. \tag{14}$$

Therefore, the estimated values of the parameters can be obtained, as shown in Table II.

According to the estimated values of the parameters, $\overrightarrow{n}$ and $\overrightarrow{R}$ in the formula cannot be obtained from the given text data, but can be inferred according to the topic $\overrightarrow{Z}$ of every word and text in the data. Therefore, Gibbs sampling is used to sample the topics from the texts or words to obtain $\overrightarrow{Z}$, and when it reaches the stationary distribution, it is the sample set we need.

C. Training Process for User Topics Mining of the PQDR-LDA-Gibbs Sampling Model

The constructed PQDR-LDA model has the following two functions.

1) With the PQDR-LDA model, data are introduced to obtain the estimated values of the model parameters $\overrightarrow{\theta_u}$, $\overrightarrow{\pi_b}$, $\overrightarrow{\pi_{ex}}$, $\overrightarrow{\phi_t}$, $\overrightarrow{\phi_b}$ and $\overrightarrow{\phi_{ex,u}}$, where $u\in[1,U]$ and $t\in[1,T]$.
2) The PQDR-LDA model with the obtained parameters can be used to mine other users U_{new} and other texts m'_{new} in the data and to obtain the topic distribution of other users $\overrightarrow{\theta_u}$, the topic of other texts $Z_{m'}$, and the additional topics of doctors' answers Z_{ex}.

By obtaining the conditional probabilities of Gibbs sampling and the estimated values of the PQDR-LDA model parameters, the PQDR-LDA model can be trained based on the data, and the trained model can be used for topic mining in new texts. The training process obtains data samples through Gibbs sampling, and all of the model parameters can be estimated based on the final samples. The training process is as follows.

1) *The Random Initialization Stage:* A disease topic Z_m is randomly assigned to every text in the data; a topic Z is randomly assigned (a disease topic Z_m or a background topic Z_b) to every word of the user's question in every text; and a topic Z is randomly assigned (a disease topic Z_m or an additional topic Z_{ex}) to every word of the doctor's answer in every text.
2) *Sampling Stage:* The data is scanned, the topic of every text is sampled according to the Gibbs sampling formula, and the topic of every word in both the user's question and doctor's answer is sampled.
3) *Repeat Stage:* The second stage is repeated through the sampling process for the data until the Gibbs sampling converges.
4) The model parameters are estimated according to the above parameter formulas.
5) The topics of every word in the data are calculated to obtain the distribution of disease topics, and the distribution of every subject word in the data is calculated to obtain the topic and word distribution of the model.

The specific Gibbs sampling method of the PQDR-LDA model is as follows:

Input: Online medical publishers' text data, including users' questions and doctors' answers (there is a one-to-one relationship between the two parts); prior parameters $\overrightarrow{\alpha}$, $\overrightarrow{\beta}$, $\overrightarrow{\beta_b}$, $\overrightarrow{\beta_{ex}}$, $\overrightarrow{\lambda_b}$, and $\overrightarrow{\lambda_{ex}}$; the number of topics T; and the frequency of iterations Q'.

Global Measurement Variables: $\overrightarrow{n_u}$, $\overrightarrow{R_b}$, $\overrightarrow{R_{ex}}$, $\overrightarrow{n_t}$, $\overrightarrow{n_b}$, and $\overrightarrow{n_{ex,u}}$.

Output: The distribution of user disease topics $\overrightarrow{\theta_u}$; the background topic correlation distribution $\overrightarrow{\pi_b}$; the additional topic correlation distribution $\overrightarrow{\pi_{ex}}$; the online medical inquiry text disease topic word distribution $\overrightarrow{\phi_t}$; the background topic word distribution $\overrightarrow{\phi_b}$; and the additional topic word distribution $\overrightarrow{\phi_{ex}}$.

IV. Experiment on Disease Topic Mining and the Result Analysis

In this section, experimental analysis is conducted to verify the effectiveness of the PQDR-LDA model. First, the preparatory work for the experiment is explained, including the analysis of experimental datasets and the model evaluation indicators. Second, the experimental results are presented and analyzed to draw experimental conclusions.

A. Experimental Dataset Analysis

Most of the data consists of the diagnosis and treatment process of user questions and doctor answers. The recorded data include physical symptoms, test results, disease symptoms, treatment conditions and other data found from the users' questions; later, doctors analyze and assess the medical conditions and then give their corresponding opinions and suggestions.

The medical entities recorded for one disease in the data may also appear in different texts. For example, common cold symptoms include headache, cough, fever, and sore throat, and the medications used to treat it include aspirin, artificial cow-bezoar, chlorphenamine maleate capsules, acetaminophen sustained-release tablets, and norfloxacin capsules. In different texts that contain the disease "cold," some or all of the physical symptoms, laboratory test results, disease symptoms, treatment conditions and other data related to a cold appear frequently. That is, the symptoms and the medications that appear in texts are disease words and medication words, respectively, and combinations of different medical condition words and medication words may appear in different texts. In addition, the order of disease words and medication words does not affect the diagnosis of the disease itself. For example, in a cold treatment text, the order of symptoms (such as headache, cough, fever, and sore throat) and the order of medications (such as aspirin, artificial cow-bezoar and chlorphenamine maleate capsules, acetaminophen sustained-release tablets, and norfloxacin capsules) will not affect the therapeutic efficacy of the disease itself. Therefore, data is characterized by strong "word order independence," which provides a prerequisite for the use of the LDA topic model for disease topic mining of texts.

In this experiment, we collected data from publishers in the "internal medicine," "surgery," "obstetrics and gynecology," "pediatrics," and "ophthalmology and otorhinolaryngology" departments of "www.120ask.com" and randomly selected more than 2000 texts from the respiratory medicine department. After data noise cleaning and preprocessing, there were a total of 1024 valid texts (which included 524 male users and 500 female users; the age range of 5–72 years old), and more than 60% of these texts were from users aged 5–45 years old [49].

The platform did not indicate the disease involved in every text from the Department of Respiratory Medicine. Therefore, to measure the experimental effect, every text was checked, the disease was determined according to the symptoms and was then manually labeled. After manual labeling, it was found that 15 common diseases, mainly influenza and respiratory tract infection, were involved in the 1024 texts. The 15 diseases involved in the texts and their approximate proportions are shown in Table III.

TABLE III
Diseases Involved in the Online Medical Inquiry Texts and the Proportions

Amygdalitis 11.65%	Gastrointestinal cold 8.45%	Influenza 7.28%
Bronchitis 6.21%	Foreign body in respiratory tract 5.32%	Pharyngitis 4.83%
Pneumonia 4.66%	Asthma 4.65%	Emphysema 4.19%
Pertussis 4.11%	Respiratory failure 3.81%	Pneumothorax 3.28%
Trachitis 3.18%	Pleurisy 2.99%	Pulmonary embolism 2.87%
Lung abscess 2.77%	Influenza Type A Subtype H1N1 0.81%	…

B. Model Evaluation Indicators

1) Perplexity: Perplexity is an indicator commonly used to measure the quality of a topic model, and it is widely used in natural language processing. Perplexity is used to measure the generalization ability of disease topic models, i.e., the ability to model new data. Generally, the smaller the perplexity value is, the stronger the generalization ability of the model, and vice versa. When the Markov chain is in the state $Q = \{\overrightarrow{W}, \overrightarrow{Z}, \overrightarrow{Y_b}, \overrightarrow{Y_{ex}}\}$, the calculation formula for the perplexity value of the test dataset D is as follows:

$$\text{Perplexity}\,(D) = p\,(\overrightarrow{w}|D) = \exp - \frac{\sum_{d=1}^{D} \log p\,(\overrightarrow{w_d}|Q)}{\sum_{d=1}^{D} N_d}. \tag{15}$$

In the formula, N_d represents all of the words in the document d; in this model, it refers to the total of the words in the users' questions and doctors' answers, and $p(\overrightarrow{w_d}|Q)$ indicates the generation probability of all words in document d. Specific to this model, the model parameters can be directly used to calculate the generation probability of document d as follows:

$$p\,(\overrightarrow{w_d}|Q) = \prod_{v=1}^{V} \left(\pi_b^1 \times \phi_b^v + \pi_b^2 \times \sum_{t=1}^{T} \theta_u^t \times \phi_t^v \right)^{(N_b^v)}$$
$$\times \prod_{v=1}^{V} \left(\pi_{ex}^1 \times \phi_{ex,u}^v + \pi_{ex}^2 \times \sum_{t=1}^{T} \theta_u^t \times \phi_t^v \right)^{(N_{ex}^v)}. \tag{16}$$

In the formula, N_b^v represents the frequency of the appearance of word v in the user's question, and N_{ex}^v represents the frequency of the appearance of word v in the doctor's answer.

2) Coherence: Coherence, as an evaluation method of probability models, is a supplement to perplexity. Related studies have suggested that coherence can be used to measure the semantic coherence of words in a topic [50]. The higher the PMI value is, the stronger the semantic coherence of the topic is, indicating that the words in the topic can support each other and that the overall performance of the topic model is better. There are many calculation formulas for semantic coherence, and the most commonly used is the C_uci method based on pointwise mutual information (PMI). Therefore, for the test data in this paper, the method was adopted as the standard for the automatic evaluation of semantic coherence within disease topics, specifically as follows.

The C_uci method is based on a sliding window to calculate the coherence of the PMI of all word pairs (one-set partitioning) under a given disease topic. The calculation formula is as follows:

$$PMI\left(w_{i''}, w_{j''}\right) = \log \frac{P\left(w_{i''}, w_{j''}\right) + \epsilon}{P\left(w_{\mathrm{i}''}\right) \cdot P\left(w_{j''}\right)}$$

$$C_{UCI} = \frac{2}{V \cdot (V-1)} \sum_{\mathrm{i}''=1}^{V-1} \sum_{j''=i''+1}^{V} PMI\left(w_{i''}, w_{j''}\right) \quad (17)$$

where $P(w_{i''})$ is the occurrence probability of word $w_{i''}$ in the test document set, $P(w_{i''}, w_{j''})$ represents the joint probability of words $w_{i''}$ and $w_{j''}$ in the test document set, and V is the dimension of the word list.

C. Experimental Results and Analysis

1) Selection and Optimization of Parameters: According to the derivation of the PQDR-LDA topic model, when the PQDR-LDA topic model is used for training and in addition to the provided data to be trained and the data to be tested, the parameters of the model should be set in advance.

a) Prior distribution parameters: In the experiment, the estimation of PQDR-LDA model parameters is based on known parameters of the Dirichlet prior distribution of the model parameters, namely, the hyperparameters α, λ_B, λ_{ex}, β, β_B, and β_{ex}. According to the nature of the Dirichlet distribution, when the parameter values are larger than 1, the probability distribution curve will be more evenly distributed, which means that the same document corresponding to the model may contain more topics; when the parameter values are less than 1, the probability distribution curve will be more concentrated, which means that there may be fewer topics in the same document corresponding to the model. This article takes the data of online medical publishers' medical inquiry texts as the object of the study. The texts were established with the users as the units. The disease topics covered among the users should be as realistic as possible, and a higher degree of distinction between disease topics means better model performance. Moreover, the Gibbs sampling algorithm requires repeated iterations. The initial values of hyperparameters α, λ_B, λ_{ex}, β, β_B, and β_{ex} have little effect on the final sampling results. Therefore, referring to the method provided by Grifliths and Steyvers [51], these hyperparameters are set as $\alpha = 50/T$, $\lambda_B = \lambda_{ex} = 0.01$ and $\beta = \beta_B = \beta_{ex} = 0.01$, and the number of Gibbs sampling iterations is 500.

b) Number of disease topics: The number of disease topics determines the distribution and quality of the PQDR-LDA model. The choice of perplexity in this article may be more helpful in selecting the number of disease topics. The smaller the perplexity value is, the better the generalization ability of the model. Under normal circumstances, perplexity will decrease as the number of disease topics increases. Therefore, we should look for the inflection point of the number of disease topics and the perplexity curve; that is, the point where the perplexity is relatively small and the number of disease topics no longer changes significantly is selected as the optimal number of topics. However, it is found that online medical entities in Chinese are not as simple as those in English, resulting in relatively large distributions of words. Therefore, it is not typically easy to find the inflection point of the number of disease topics and the perplexity curve. With the increase in the number of disease topics, the perplexity of the PQDR-LDA model increase or decrease in texts (involving diseases of the gastroenterology department, nephrology department and others). Therefore, introducing model coherence will be more helpful for the selection of the number of disease topics.

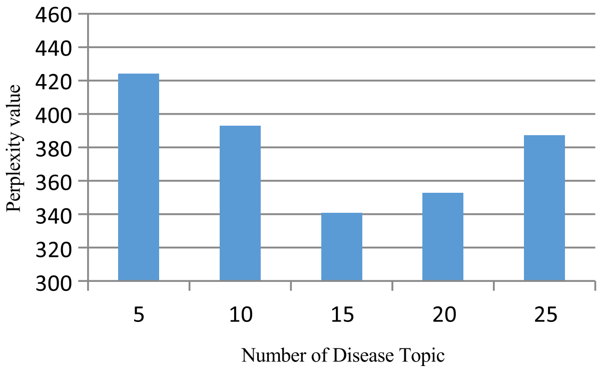


Fig. 2. Changes in the disease topic perplexity value of the PQDR-LDA model under the training set.

According to the characteristics of the wholly obtained 1024 texts released by medical publishers in the Department of Respiratory Medicine, in this article, the number of disease topics was set between 5 and 25. Then, the final number of disease topics was selected according to the perplexity value. The experimental results are shown in Fig. 2, where it can be seen that when $T$is 15, the model has a minimal perplexity value. This means that when the number of disease topics is 15, the degree of distinction between disease topics is higher, the coherence between words within every disease topic is higher, and the overall model is optimal.

As previously mentioned, in this article, the number of disease topics was set between 5 and 25. Then, the final number of disease topics was selected according to the coherence value. The experimental results are shown in Fig. 3, where it can be seen that when $T$is 15, the model has a minimal coherence value. This means that when the number of disease topics is 15, the model has the lowest perplexity value and the highest coherence degree, suggesting the optical quality of the model. Therefore, in the training set formed by the obtained 1024 texts released by the medical publishers in the Department of Respiratory Medicine, the number of disease topics was 15.

Under the training set formed by the wholly obtained data of the 1024 texts released by the medical publishers in the Department of Respiratory Medicine, all of the online medical inquiry texts of the user "Uid = 7016" were used as the basis. In the experiment, we utilized the result of the superparameter setting of the PQDR-LDA model and adopted the perplexity

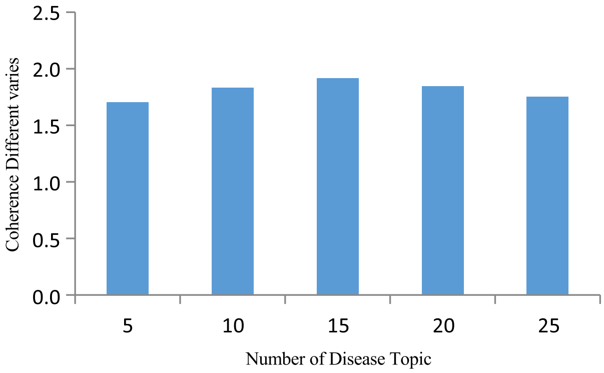


Fig. 3. Changes in the disease topic coherence value of the PQDR-LDA model under the training set.

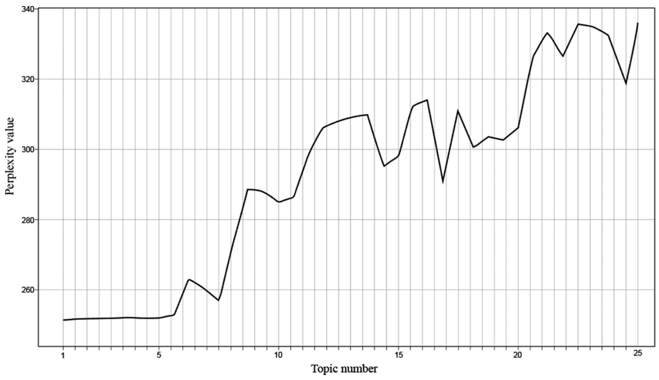


Fig. 4. Uid = 7016 user-based perplexity values of the PQDR-LDA model with different numbers of disease topics.

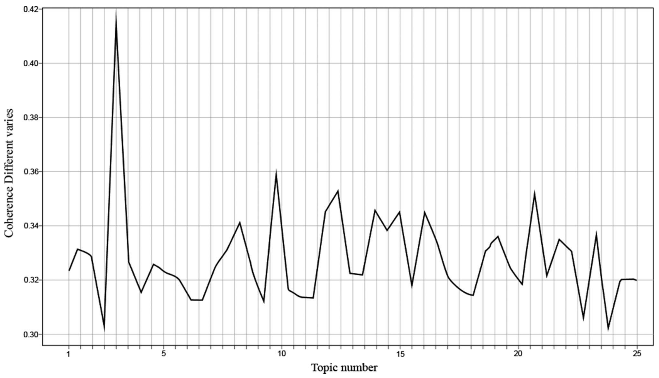


Fig. 5. Uid = 7016 user-based coherence values of the PQDR-LDA model with different numbers of disease topics.

method to evaluate the quality of the model. With the number of disease topics T as the abscissa and the perplexity value of the model as the ordinate, we drew a broken line graph, as shown in Fig. 4.

Within the scope $T \in [1, 25]$, the overall perplexity level of the model is relatively low, which also verifies that when user's data are merged into a document, the PQDR-LDA model for texts based on the unit of user shows better performance. Taking a closer look, when $T < 5$, the model has relatively low perplexity values, but there is little change and no obvious inflection point; when $T > 5$, the perplexity of the model generally increases with the number of disease topics. Based on the line graph, we could only narrow down the scope of the number of disease topics T to $[1, 5]$ and could not obtain a more specific optimal number of topics.

Based on these results, the coherence method can be used to further determine the optimal number of disease topics in the model. With the number of disease topics T as the abscissa and the coherence value of the model as the ordinate, we drew a broken line graph, as shown in Fig. 5.

Similar to the previous examination, all of the texts of user "Uid = 7016" were used as a basis. Within the range $T \in [1, 25]$, the model coherence fluctuates and has a maximum value at the peak when $T = 3$. This means that when the number of disease topics is 3, the degree of distinction between disease topics is higher, the interword coherence within every disease topic is higher, and the overall model is optimal. In addition, $T = 3$ also falls in the optimal perplexity range. For this case of PQDR-LDA modeling, the optimal number of user disease topics is 3, as the arrangement has the lowest perplexity value, the highest coherence, and the best overall quality of the model.

2) Training and Testing of the PQDR-LDA Model: This article uses the PQDR-LDA model to mine information that can describe users' disease topics from the data released by online medical publishers and to obtain a set of word distributions used to describe the disease topics texts.

a) Disease topic-lexical item distribution in the training set: In this section, the training set formed by the wholly obtained data of the 1024 texts released by the medical publishers in the Department of Respiratory Medicine was used as the basis to train the PQDR-LDA model. The model gathered online medical inquiry texts with the user as the unit. Referring to the conclusion drawn in Section IV-C1, the hyperparameters were set as follows: α as 50/T, β as 0.01, the number of Gibbs sampling iterations as 500, and the optimal number of topics in the model as 15. Then, we calculated the key parameters $\overrightarrow{\theta_u}$ and $\overrightarrow{\varphi_t}$ of the model.

Although the Department of Respiratory Medicine under the online medical platform has also classified the disease labels (with a total of 8), it does not mean that the optimal number of disease topics in the model is 8. Due to the coarse granularity of the classification labels of the respiratory diseases under the platform, many disease classification labels belong to different categories (such as "trachitis," "cold," and "pharyngitis"). The words in these classification labels can be further classified in ICD-10. After the PQDR-LDA topic model training, different topics will be subdivided. In this sense, the optimal number of disease topics for the model is 15, which is greater than the number of classification labels. Through a multidimensional scaling (MDS) algorithm [52], the disease topic-word high-dimensional probability matrix of the training set composed of parameter $\overrightarrow{\varphi_t}$ can be visualized in a two-dimensional (2-D) space. The 2-D results of the algorithm also show better distinction between the disease topics of the model when the number of topics is 15.

As shown in Fig. 6 , when the number of disease topics in the left panel is 15, there is no intersection between the

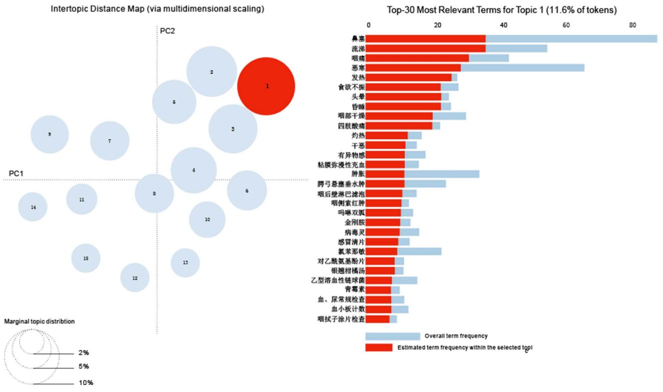


Fig. 6. MDS algorithm result of the training set of the topic lexical item high-dimensional probability matrix.

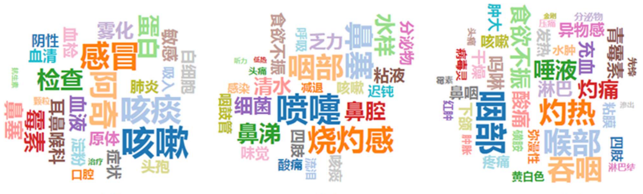


Fig. 7. Word cloud graph of the disease topic lexical item distribution of user "Uid = 1927." (a) Distribution of words under topic 1. (b) Distribution of words under topic 2. (c) Distribution of words under topic 3.

TABLE IV
ARTIFICIAL LABELS OF THE DISEASE TOPICS IN THE TRAINING SET MODEL

Topic No.	Artificial Label	Topic No.	Artificial Label
1	Tonsillitis	9	Emphysema
2	Gastrointestinal Cold	10	Pertussis
3	Influenza	11	Respiratory failure
4	Bronchitis	12	Pneumothorax
5	Foreign body in the respiratory tract	13	Trachitis
6	Pharyngitis	14	Pleurisy
7	Pneumonia	15	Lung abscess
8	Asthma		

circles of disease topic-lexical item distribution, suggesting a separation between disease topic-lexical item distribution and good distinction between disease topics in high dimensions, so the model has a better training effect on the training set.

Based on Section IV-B, combined with the lexical item distribution in every disease topic and the corresponding word cloud graph, we artificially labeled the 15 topics of the trained model, and the results are shown in Table IV.

The essence of using the trained model for inference is to estimate the parameter $\overrightarrow{\theta_u}$ of an unknown document based on the known parameter $\overrightarrow{\varphi_t}$ with a better distinction degree. Therefore, the information of the trained model parameter $\overrightarrow{\varphi_t}$ should be saved in a specific form for later inference. The overall effect of the PQDR-LDA model on disease topic-lexical item distribution is shown in Tables V and VI.

Due to space considerations, only the high-frequency words of 3 of the disease topics are listed. Based on the words corresponding to the disease topics, it can be seen that the disease topics indicated by $topic1$, $topic2$, and $topic3$ should be amygdalitis, gastrointestinal cold, and influenza, respectively. It is found that the three disease topics have clear definitions, they are all differentiable and the words in the disease topics are also highly related. The PQDR-LDA model can effectively learn the disease topics in the online medical publishers' medical inquiry text data.

TABLE V
DISEASE TOPIC-LEXICAL ITEM DISTRIBUTION LIST EXTRACTED WITH THE PQDR-LDA MODEL FROM THE ONLINE MEDICAL INQUIRY TEXTS OF THE RESPIRATORY MEDICINE DEPARTMENT

topic 1	Nasal obstruction (0.044880965), runny nose (0.044841402), sore throat (0.044258343), chills (0.040336513), fever (0.038262528), loss of appetite (0.036433022), dizziness (0.036214188), lethargy (0.034136519), dry throat (0.033430353), , , sore limbs (0.031189793), burning (0.032120523), vomiturition (0.030568820), foreign body sensation (0.029655821), diffuse mucosal hyperemia (0.028883101), swelling (0.025777723), palatal arch staphylgroma (0.022792166), posterior pharyngeal wall lymphoid follicles (0.021531069), lateral pharyngeal redness and swelling (0.020199044), mandibular lymphadenopathy (0.017288329), sulfa drugs (0.016829175), cephalosporin (0.015885636), morpholine biguanide (0.015412230), amantadine (0.014040804), , , varustat (0.011710238), Ganmaoqing tablets (0.011508385), chlorpheniramine (0.009442326), acetaminophen tablet (0.008215502), Yinqiao Ganju decoction (0.007260846), beta hemolytic Streptococcus (0.005592498), penicillin (0.003686038), routine blood and urine tests (0.002264307), platelet count (0.002083467), pharyngeal swab smear (0.002081625)······
topic 2	abdominal pain (0.039091334), diarrhea (0.039131298), vomiting (0.038680893), running nose (0.030433909), nasal congestion (0.029131298), stomach cramps (0.028932746), aversion to cold (0.028369710), Huoxiang Zhengqi liquid (0.027828688), Fuke'an (0.025812540), anisodamine (0.024952381), norfloxacin (0.022692347), metoclopramide montmorillonite (0.02037892), ciprofloxacin (0.019223468), erythromycin ethylsuccinate (0.015223468), ribavirin (0.009323268), dexamethasone (0.005578234)······
topic 3	sore throat (0.038892877), dysphagia (0.03777002), fever (0.036050735), cough (0.024173591), limb muscle soreness (0.011027399), throat redness and swelling (0.009454304), cephalexin (0.007881208), Pudilan capsule (0.007532562), Qingresanjie tablet (0.006308113), nasal mucosal cell test (0.006307982), serum internal antibody test (0.004735017), moroxydine hydrochloride (0.004624026), whole virion inactivated vaccine (0.004583358)······

Due to the certain interdisease relevance, additional topics from doctors' potential supplementary diagnosis suggestions for every user can serve as a supplementary basis to obtain a precise user's disease topic. It can be seen from Table VI that additional topics have a certain meaning that relates user topics. According to the words that correspond to the additional topics, it can be seen that the accompanying disease topics for User 1, User 2, and User 3 should be bronchitis, acute gastroenteritis, and pneumonia, respectively. In addition, the additional topics also contain some words related to the disease topics of the texts; obtaining the distribution of these words can reflect the accompanying diseases that might be induced by the major diseases of the users in reality and thus better assess the current disease conditions of the users.

b) Disease topic-lexical item distribution under the test set: In this section, all of the online medical inquiry texts of user "Uid = 1927" in the test set formed from the globally obtained data of the 1024 medical inquiry texts released by the online

TABLE VI
ADDITIONAL TOPIC-LEXICAL ITEM DISTRIBUTION LIST MINED WITH THE PQDR-LDA MODEL

User 1	cough (0.019382945), expectoration (0.017202162), asthma (0.015528539), fever (0.015274959), sore throat (0.009753483), antibiotics such as cephalexin (0.009702767), ambroxol (0.009601335), carbocysteine (0.008449187), theophylline sustained-release tablet (0.008163512), compound liquorice tablet (0.007688449), compound methoxyphenamine (0.006637733), montelukast sodium tablet (0.004485585), bronchography (0.003384153), X-ray examination (0.00234367)······
User 2	Redness and swelling (0.01792526), tenderness (0.01598797), cramps (0.01356623), stiffness (0.00872294), muscle tension (0.00823852), oral glucose (0.00726992), intramuscular injection of chlorpromazine (0.00678554), atropine sulfate (0.00630125), dioctahedral smectite (0.00630126), berberine (0.00533253), routine stool examination (0.00484821), white blood cell count (0.00436393), diclofenac sodium enteric-coated tablet (0.00434964), ibuprofen (0.00387955), celecoxib (0.00339526), glucosamine (0.00223461), meloxicam (0.00133252)······
User 3	headache (0.019789778), sweating (0.015103952), polypnea (0.011980074), high fever (0.011459435), shortness of breath (0.01143396), purulent sputum (0.00833527), chest pain (0.007294263), Maxingshigan decoction (0.007283665), Sanhuangshi decoction (0.005211677), cephalosporin (0.004691021), macrolide antibiotics (0.004670123), ambroxol hydrochloride (0.004551025), Chuanbei Pipa cough syrup (0.004391026), montmorillonite powder (0.003129098), chest X-ray examination (0.003129094), blood examination (0.00313973), sputum cell culture (0.00309526)······

medical publishers in the Department of Respiratory Medicine were used as the basis, and we used the PQDR-LDA model for the modeling. According to the conclusions drawn in Section IV-C1, online consultation texts were collected on a user basis, and the hyperparameters were set accordingly, α as 50/T, β as 0.01, the number of Gibbs sampling iterations as 500, and the optimal number of model topics as 3. According to the training model, we obtained $\overrightarrow{\phi_t}$ and calculated the key parameter of the model $\overrightarrow{\theta_u}$.

In the Fig. 7, topic 1 can be classified as a "pneumonia" theme; similarly, topics 2 and 3 can be classified as having themes of "influenza" and "pharyngitis," respectively.

The MDS algorithm is a data dimensionality reduction method similar to principal component analysis. It utilizes the pairwise similarity of samples to construct a relatively low-dimensional space, making the distance of every pair in the high-dimensional space as consistent as possible with the sample distance in the constructed low-dimensional space and therefore showing the characteristics of high-dimensional multivariate data in the low-dimensional space. Furthermore, through the MDS algorithm, the user disease topic-word high-dimensional probability matrix composed of parameter $\overrightarrow{\varphi_t}$ is visualized in a 2-D space. The 2-D results of the algorithm are shown in Fig. 8.

As shown in Fig. 8, the circles on the left panel represent the distribution of different disease topics and the distances between them. Similar disease topics are closer together, while different disease topics are farther apart. The relative size of the distribution circle of the disease topic lexical item corresponds to the relative probability of the disease topic in a certain user's online medical inquiry texts. The three circles represent that the distributions of the three disease topics, instead of overlapping each other in the 2-D scope, are far apart, indicating good high-dimensional distinction between the disease topics. At this time, the model works well on training certain users' online medical inquiry texts. The right panel shows the probability of every word generated under the topic of the red circle. On the graph, the words "cough," "cold," and "sputum" are more likely to be generated, which is consistent with the given word cloud.

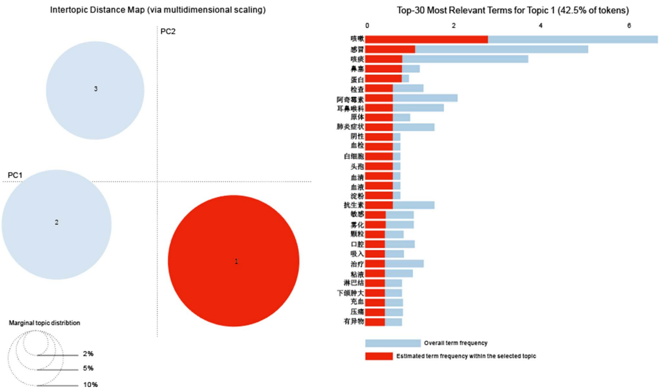


Fig. 8. MDS algorithm result of the Uid = 1927 user's disease topic lexical item high-dimensional probability matrix.

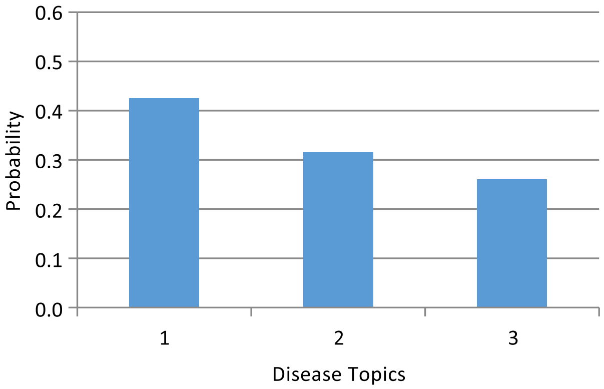


Fig. 9. Distribution of the Uid = 1927 user's disease topics.

The model parameter $\overrightarrow{\theta_u}$ can represent the user-topic distribution, and each component value θ_u represents the probability of generating a certain disease topic from the user text. Since the words of the user's questions and doctor's answers were merged in every online medical inquiry text of the user and distributed during the research to determine the word directors, only the online medical inquiry text-topic distribution is displayed here, and it is shown as a histogram in Fig. 9.

When analyzing the distribution of user's disease topic lexical items, we manually added classifications for the three topics. As shown in Fig. 9 above, in the online medical inquiry texts on a user basis, the generation probability of Topic 1 "pneumonia" exceeds 40%, and those of Topic 2 "influenza" and Topic 3 "pharyngitis" exceed 30% and 25%, respectively. It can be considered that the user mainly asked about pneumonia-based diseases in the online medical inquiry platform.

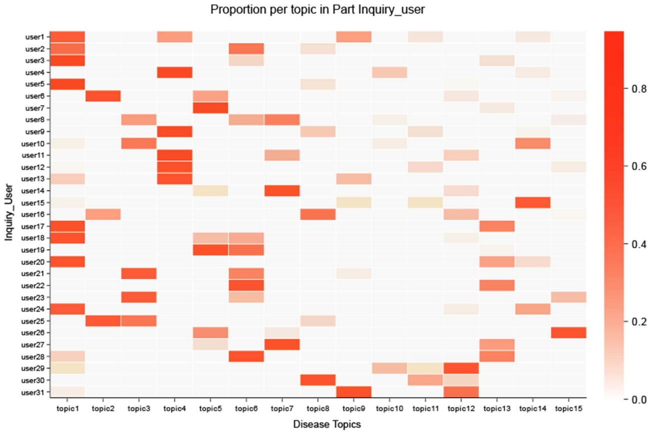


Fig. 10. Thermodynamic diagram of some users' disease topics distributions.

In summary, based on the word cloud graph and histogram, it can be considered that the online medical inquiry text data of user "Uid = 1927" reflect that their most concerning topic is "pneumonia," and the user also used the online medical inquiry platform to research "influenza" and "pharyngitis."

We used the same method to analyze all 31 users' online medical inquiry texts in the data of the medical inquiry texts released by the online medical publishers in the test set and applied the PQDR-LDA model trained with the training set for analysis. We directly used the parameter setting of the trained PQDR-LDA for the other information, including hyperparameters, optimal number of topics in the model, and key parameter $\overrightarrow{\varphi_t}$, and calculated the key parameter $\overrightarrow{\theta_u}$ of the model.

The parameter $\overrightarrow{\theta_u}$ represents the user-topic distribution, and every component value θ_u represents the generation probability of a certain disease topic under the user. In this section, a total of 31 user-topic distributions were randomly selected and visualized in a thermodynamic diagram with the number of topics as the abscissa and the user number as the ordinate.

As shown in Fig. 10 above, the darker the grid color, as displayed on the right panel of the thermodynamic diagram, the greater the probability of occurrence of the disease topic in the corresponding user's online medical inquiry texts. Then, through the thermodynamic diagram, the distribution of every disease topic by every user (taken as a base unit) can be clearly seen. Among them, every disease topic accounted for an even proportion, suggesting that the 31 users followed specific contents of the disease topics 1–15 in the online medical inquiry texts.

3) *Comparative Analysis of the PQDR-LDA Model:*

a) Perplexity value comparison: First, to study the impact of changes in the number of disease topics T on the two models, T was set to different values for the experiment. In addition, to shield the influence of iterations, the number of iterations is set to a fixed value, 500. The obtained perplexity values of the two models are shown in Fig. 11.

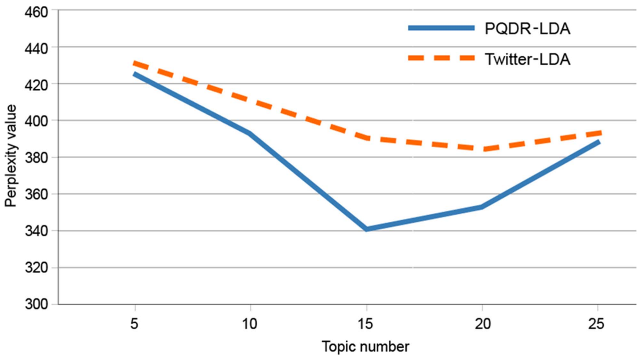


Fig. 11. Comparison of perplexity value between the PQDR-LDA model and the twitter-LDA model under different T values.

Fig. 11 shows that the curves of the both models show a trend of first decreasing and then increasing; when the number of disease topics T is at a certain value, the perplexity is the smallest. As described in the previous section, the smaller the perplexity value is, the stronger the generalization ability of the model, so the generalization abilities of the models first rise and then decrease with the number of disease topics. This result can be easily understood. The number of disease topics in specific text data is objectively fixed, and defining too many or too few disease topics from the dataset weakens the effect of the models. In addition, the PQDR-LDA model achieves an optimal performance when the number of disease topics $T = 15$, while the Twitter-LDA model is optimal when the number of disease topics $T = 20$. These results are expected, as while the PQDR-LDA model can mine additional topics from different users, the Twitter-LDA model can only mine background topics shared by every user and present the additional topics related to the users in the form of average disease topics. In this sense, the latter works better with a larger number of disease topics, and the extra disease topics (compared with the case in the former model) are likely to be additional topics related to the users. In addition, with the number of disease topics between 5 and 25, the whole perplexity curve of the PQDR-LDA model is below that of the Twitter-LDA model; therefore, the PQDR-LDA model proposed in this paper has a stronger generalization ability.

Then, the influence of model iterations on the model performance was explored. Similarly, to shield the influence of the number of disease topics, a fixed value was set, and the number of model iterations was changed to obtain the corresponding perplexity values. Since it has been suggested by the results of the previous experiment that the PQDR-LDA and Twitter-LDA models work best when the numbers of disease topics are 15 and 20, respectively, the numbers of disease topics were set to the corresponding optimal values for the two models. The obtained perplexity values of the two models are shown in Fig. 12.

It can be seen from the above figure that the perplexity values of the two models show a decreasing trend, with the decrease becoming increasingly smaller as the number of iterations increases. The number of iterations represents the number of model *Gibbs* samplings. As the number of model samples increase, the Markov chain tends to converge, and each distribution in the model becomes closer to the true probability. Theoretically, a large number of iterations should be set for the models. However, in practice, due to resource and efficiency considerations, a compromised number of iterations is usually set, which can lead to a better model effect without consuming

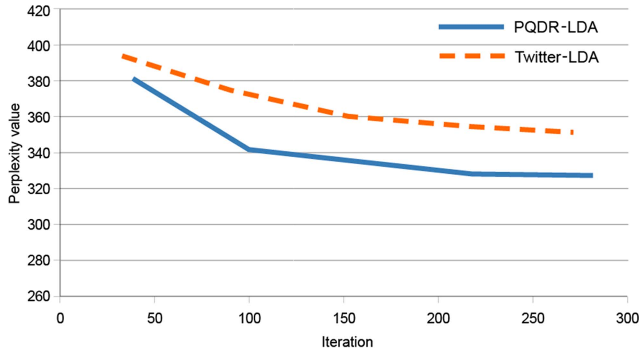


Fig. 12. Comparison of perplexity value between the PQDR-LDA model with $T = 15$ and the twitter-LDA model with $T = 20$.

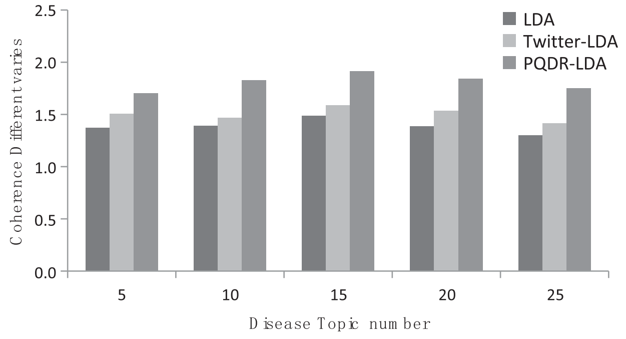


Fig. 13. C_uci values of the disease topics in the test dataset.

too many resources. For example, the number of iterations set in this experiment (250) is a good choice. In addition, similar to the result of the previous experiment, the curve of the PQDR-LDA model is also below that of the Twitter-LDA model, indicating that the former has stronger generalization ability.

b) Coherence value comparison: Fig. 13 shows the C_uci values of the three topic models (LDA, Twitter-LDA, and PQDR-LDA) under different numbers of disease topics T.

It can be seen from Fig. 13 that with the test dataset, the C_uci values generated by the PQDR-LDA model proposed in this paper are generally higher, indicating stronger semantic coherence in the extracted disease topic words. Applied to online medical inquiry texts, the traditional LDA model does not take into account the abundance of unstructured data and the complexity of words. The content cannot fully reflect the online medical entities followed by users, resulting in insufficient semantic enhancement of the texts and words. Therefore, the model has the lowest C_uci values. The Twitter-LDA is an extended structure of the LDA model. Due to the introduction of background topics, irrelevant narrative characters are removed, and most of the remaining content contains online medical entities. Therefore, the semantics of the words have been strengthened, and the Twitter-LDA has higher C_uci values than the LDA model. However, the Twitter-LDA model does not merge the medical entities in the users' question and those in the doctor's answers to strengthen the semantics of the texts and words, and there is only random fusion of the content. The PQDR-LDA disease topic model proposed in this article has even higher C_uci values. This is because in essence, the medical entities in the users' questions are characterized by sparse medical entities and text semantic fuzziness, while the doctors' answers have dense medical entities and text semantic clearness. Therefore, instead of ordinary content fusion, the PQDR-LDA model further merges the specific medical entity words in the users' questions and the doctors' answers in the texts, thereby forming a more efficient probability model, which is more in line with the probability calculation mode of the C_uci value.

c) Result and analysis: Based on the above experiments, it can be seen that compared with the Twitter-LDA model, the PQDR-LDA model has better performances in regard to both perplexity and coherence values. The reasons are as follows.

1) Using the Twitter-LDA model to mine users' questions alone might not be able to precisely acquire the disease symptoms expressed by the users and mining the doctors' answers alone sometimes fails to capture users' health claims. The PQDR-LDA model can make up for these problems, as it expands the content of the users' questions from the content of the doctors' answers to precisely mine the disease topics. Therefore, the PQDR-LDA model has a better mining ability.
2) Another difference between the Twitter-LDA model and the PQDR-LDA model is that under the premise of retaining background topics, the PQDR-LDA model sets an additional topic for each user; the additional topics of different users have different word distributions. This contributes to the higher mining precision of the model.
3) In most cases, the doctor's answer shares the disease topic related to the user's question. However, the concurrency and relevance of diseases lead to many situations in which doctors' answers share the same disease topics followed by users' questions or doctors' answers have disease topics related to those followed by users' answers. If this part of the content is merged with the users' question content and mined with the Twitter-LDA model, then user-related disease topic content might be mistaken by the model as being independent of the user's question and thus may be discarded. The PQDR-LDA model can be used to mine user-related content, thereby improving the performance of user disease topic mining.

V. Discussion and Conclusion

Although the consultation data on online medical platform can produce great benefits for online medical operation service, the platform operator needs to take into account the difference in perception of medical concepts and health information between medical professionals and users in their consultation services. In this study, we focus on the cognitive difference of consultation service in platform operations, which is called disease topic. Consultation-based services are common on online healthcare platforms. Cognition of diseases based on consultation services is a common way to obtain medical information, specifically, users obtain medical information from consultation texts of online medical platforms to help themselves recognize diseases. However, the nature of users' cognition of disease poses unique challenges to platform operators' service strategies in operations. With the increase of users' attention to their own diseases, due to the sparse and dense medical entities in the consultation service texts, as well as the ambiguity and clarity of the text semantics, mismatching can easily occur in mining the disease topic, causing distortion to user's cognitive disease, lowering the quality of consultation services, and eventually leading to the loss of revenue and user base. On the one hand, we need to obtain higher quality disease topics in multiple consultation text questions and multiple doctors' answers; on the other hand, the consultation environment can create stronger cognitive differences, which is more likely to cause disease topic misalignment, lower demand and affect profitability. Therefore, platform operators should optimally balance these two aspects when formulating service strategies.

We establish a topic model PQDR-LDA based on the association between users' questions and doctors' answers. Doctor's answer text is considered to act as the context of users' question text under obscure subject, so as to make the subject of users' question text clear. We set additional topics in doctors' answer text as a supplementary basis for users' disease topics, and different users have different additional topics, hence, we further dig out user-specific concurrent disease topics. This not only reduces the impact on users' cognitive difference of disease topics in multiconsultation text questions and multiple doctors' answers, so as not to cause disease topic misalignment, but also obtains other disease topics hidden by users from doctors' answer text.

A. Managerial Implications

In this study, we specifically attempt to answer some important questions with significant management implications to both platform operators and users in online medical consultation services. Our first question is: can the use of a single consultation text for raising question and a single doctor's answers in online medical consultation text reduce the misalignment of disease topic mining, thereby reducing user loss, and ultimately optimizing the service strategy of platform operators? We found that this is not necessarily true. In fact, the existing scenario of single question and single answer exacerbates the occurrence of disease topic mining misalignment. This result has important management implications, because a single question and a single answer has recently become an increasingly common phenomenon in order to conserve medical resources. Our results suggest that using the model in this paper to mine disease topics in multiconsultation text questions and multiple doctors' answers can avoid misalignment, thereby improving the profitability of online medical consultation services and enhancing the brand image of platform. Platform operators need to balance the relationship among profitability, image, resources and user stickiness.

Regarding the second research question, we expected to figure out whether lower technology cost can always benefit online healthcare platform operators. Contrary to universal opinions, our results suggest that even if the cost of technology becomes lower for online healthcare platforms, the reality of platform operators may be worse. This is an interesting finding, and has considerable implications for platform operators. Because the consultation users on online medical platforms are more sensitive to the results of disease topic mining, they (platform operators) must carefully evaluate the pros and cons of technology costs in this environment, so as to determine whether to expand market competition by adopting technological innovation (model expansion), rather than through price reduction or other service strategies. It is worth noting that reducing technical costs and improving the capacity of platform knowledge base are not necessarily beneficial to platform consultation users.

B. Theoretical Implications

In this study, for platform operators and users involved in online medical consultation services, we try to answer important questions that can expand theoretical innovation. Our first question is: what happens to platform consultation users when platform operators' technological innovation (model expansion) leads to increased costs, or more formally speaking, when technology becomes more expensive for platform operators, do platform operators transfer the cost to platform consultation users? With the increase in the cost of technology, online medical platform operators usually do not transfer the cost to platform consultation users at the current stage, especially when the cost of switching medical platforms is low. However, it can be observed from online medical platforms that not or rarely transferring cost is not always the optimal strategy. In some cases, platform operators can actually transfer cost to platform consultation users in a rhythmic and hierarchical manner. This result clarifies the optimal profit distribution arrangement between platform operators and platform consultation users, and provides useful insights for platform operators. This result reveals the need for optimal profit distribution between platform operators and users. Since it is the first time to analyze the cost strategy transformation brought about by technological innovations (model expansion) in online medical scenarios, and the research of "cost transfer theory" by previous scholars has been extended, this study has important theoretical expansion value.

The second question is, when platform consultation users become more intolerant of disease topic mining misalignment, should platform operators offer higher discounts to attract these consultation users? We found that when platform consultation users are more sensitive to misalignment of disease topic mining, platform operators can provide technological innovation (model expansion) for more sensitive consultation users to give them better services, so as to raise prices, instead of blindly reducing prices without categorizing consultation users. This result somewhat goes against our intuitive, because the observation of examples shows that when the quality of consultation services declines due to misalignment of disease topic mining, platform operators tend to offer higher discounts, which is actually not a wise strategy. At the same time, we also noticed that when platform consultation users are more sensitive to misalignment of disease topic mining, the total demand for online medical platform consultation services may rise. This result shows that platform operators and platform consultation users need to achieve a balance among discounts, services and prices, especially in the case of a rise in the aggregate demand. For the first time, the service differences brought by technological innovations (model expansion) in online medical scenarios are included in the pricing strategy, which makes up for the gap in academic research of differentiated services of online medical platforms, and deepens the research of "differential pricing theory" in the domain of health economics.

C. Limitations and Future Research Directions

Future research will further explore this topic. First, this model makes use of the information from users' questions and doctors' answers. In fact, other texts related to user texts can also have a good revealing effect on disease topic mining, and they can be combined with users' online inquiry texts in future research. Second, this model treats the impact of each online inquiry text on a certain disease topic of users equally. However, the quality of multiple online inquiry texts of the same user varies greatly. In future studies, the influence of doctors' qualification and experience on a certain disease topic of users can be considered, and then the contribution weight of each online inquiry text to a certain disease topic of users can also be considered. Finally, this model only considers the influence of users' question texts and doctors' answer texts on the disease topic in online medical inquiry, but it fails to include other influencing factors. In future studies, the sentiment polarity factor can be added to the mining of texts, and the users concerns can be enhanced by using the information of the influencing factor.

REFERENCES

[1] S Federico, "These three sectors will benefit the most in the future from the growing size of China's digital healthcare market," 2017, Accessed: Nov. 14, 2021. [Online]. Available: https://www.sohu.com/a/130480251_116132

[2] J. Khuntia, S. Mithas, and R. Agarwal, "How service offerings and operational maturity influence the viability of health information exchanges," *Prod. Operations Manage.*, vol. 26, no. 11, pp. 1989–2005, 2017.

[3] D. H. Saifee, I. R. Bardhan, A. Lahiri, and Z. Zheng, "Adherence to clinical guidelines, electronic health record use, and online reviews," *J. Manage. Inf. Syst.*, vol. 36, no. 4, pp. 1071–1104, 2019.

[4] T. H. Hejazi, H. Badri, and K. Yang, "A reliability-based approach for performance optimization of service industries: An application to healthcare systems," *Eur. J. Oper. Res.*, vol. 273, no. 3, pp. 1016–1025, 2019.

[5] H. Giunt, "Treatment incentives and the nature of the doctor–patient relationship," *Amer. J. Bioeth.*, vol. 16, no. 10, pp. 77–78, 2016.

[6] Y. A. Qadri, A. Nauman, Y. B. Zikria, A. V. Vasilakos, and S. W. Kim, "The future of healthcare internet of things: A survey of emerging technologies," *IEEE Commun. Surv. Tut.*, vol. 22, no. 2, pp. 1121–1167, Apr.–Jun. 2020.

[7] T. R. Goodwin and S. M. Harabagiu, "Knowledge representations and inference techniques for medical question answering," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 2, pp. 14.1–14.26, 2017.

[8] Y. Li et al., "Extracting medical knowledge from crowdsourced question answering website," *IEEE Trans. Big Data*, vol. 6, no. 2, pp. 309–321, Jun. 2020.

[9] J. Jin, X. Yan, Y. Li, and Y. Li, "How users adopt healthcare information: An empirical study of an online Q&A community," *Int. J. Med. Informat.*, vol. 86, pp. 91–103, 2016.

[10] D. Wani and M. Malhotra, "Does the meaningful use of electronic health records improve patient outcomes?," *J. Operations Manage.*, vol. 60, pp. 1–18, 2018.

[11] X. Liu, Y. Zhou, and W. Zongrun, "Can the development of a patient's condition be predicted through intelligent inquiry under the e-health business mode? Sequential feature map-based disease risk prediction upon features selected from cognitive diagnosis big data," *Int. J. Inf. Manage.*, vol. 50, no. 2, pp. 463–486, 2020.

[12] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *Eur. J. Oper. Res.*, vol. 267, no. 2, pp. 687–699, 2018.

[13] W. B. Arfi, I. B. Nasr, G. Kondrateva, and L. Hikkerova, "The role of trust in intention to use the IoT in eHealth: Application of the modified UTAUT in a consumer context," *Technological Forecasting Social Change*, vol. 167, no. 3, pp. 23–40, 2021.

[14] A. Adé et al., "Chronic patients' satisfaction and priorities regarding medical care, information and services and quality of life: A French online patient community survey," *BMC Health Serv. Res.*, vol. 20, no. 1, pp. 87–99, 2020.

[15] R. Mousavi, T. S. Raghu, and K. Frey, "Harnessing artificial intelligence to improve the quality of answers in online question-answering health forums," *J. Manage. Inf. Syst.*, vol. 37, no. 4, pp. 1073–1098, 2020.

[16] Y. Yang, X. Zhang, and P. K. C. Lee, "Improving the effectiveness of online healthcare platforms: An empirical study with multi-period patient-doctor consultation data," *Int. J. Prod. Econ.*, vol. 207, pp. 70–80, 2019.

[17] J. Marynissen and E. Demeulemeester, "Literature review on multi-appointment scheduling problems in hospitals," *Eur. J. Oper. Res.*, vol. 272, no. 2, pp. 407–419, 2019.

[18] A. J. Lazard, E. Scheinfeld, J. M. Bernhardt, G. B. Wilcox, and M. Suran, "Detecting themes of public concern: A text mining analysis of the centers for disease control and prevention's ebola live twitter chat," *Amer. J. Infection Control*, vol. 43, no. 10, pp. 1109–1111, 2015.

[19] L. Yan and Y. Tan, "The consensus effect in online health-care communities," *J. Manage. Inf. Syst.*, vol. 34, no. 1, pp. 11–39, 2017.

[20] S. Bar-Lev, ""We are here to give you emotional support": Performing emotions in an online HIV/AIDS support group," *Qualitative Health Res.*, vol. 18, no. 4, pp. 509–521, 2008.

[21] K. Y. Chuang and C. C. Yang, "Informational support exchanges using different computer-mediated communication formats in a social media alcoholism community," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 1, pp. 37–52, 2014.

[22] P. Biyani, C. Caragea, P. Mitra, and J. Yen, "Identifying emotional and informational support in online health communities," in *Proc. COLING 25th Int. Conf. Comput. Linguistics*, 2014, pp. 827–836.

[23] P. K. H. Mo and N. S. Coulson, "Exploring the communication of social support within virtual communities: A content analysis of messages posted to an online hiv/aids support group," *CyberPsychol. Behav.*, vol. 11, no. 3, pp. 371–374, 2008.

[24] C. A. Heidelberger, O. El-Gayar, and S. Sarnikar, "Online health Social networks and patient health decision behavior: A research agenda," in *Proc. IEEE Hawaii Int. Conf. Syst. Sci.*, 2011:pp. 1–7.

[25] H. Christensen, K. Griffiths, C. Groves, and A. Korten, "Free range users and one hit wonders: Community users of an internet-based cognitive behaviour therapy program," *Australian New Zealand J. Psychiatry*, vol. 40, no. 1, pp. 59–62, 2010.

[26] C. F. V. Uden-Kraan, C. H. C. Drossaert, E. Taal, E. R. Seydel, and M.A.F.J. van de Laar, "Participation in online patient support groups endorses patients' empowerment," *Patient Educ. Counseling*, vol. 74, no. 1, pp. 61–69, 2009.

[27] P. Selby, T. V. Mierlo, S. C. Voci, D. Parent, and J. A. Cunningham, "Online social and professional support for smokers trying to quit: An exploration of first time posts from 2562 members," *J. Med. Internet Res.*, vol. 12, no. 3, 2010, Art. no. e34.

[28] X. Tang and C. C. Yang, "Ranking user influence in healthcare social media," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 1–21, 2012.

[29] P. K. H. Mo and N. S. Coulson, "Developing a model for online support group use, empowering processes and psychosocial outcomes for individuals living with HIV/AIDS," *Psychol. Health*, vol. 27, no. 4, pp. 445–459, 2012.

[30] C. Bao, I. R. Bardhan, H. Singh, B. A. Meyer, and K. Kirksey, "Patient-provider engagement and its impact on health outcomes: A longitudinal study of patient portal use," *MIS Quart., Manage. Inf. Syst.*, vol. 44, no. 2, pp. 699–723, 2020.

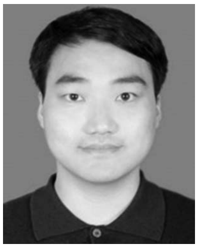
[31] M. Burri, V. Baujard, and J.-F. Etter, "A qualitative analysis of an internet discussion forum for recent ex-smokers," *Nicotine Tobacco Res.*, vol. 8, pp. S13–S19, 2006.

[32] X. Luo, B. Gu, J. Zhang, and C. W. Phang, "Expert blogs and consumer perceptions of competing brands," *MIS Quart., Manage. Inf. Syst.*, vol. 41, no. 2, pp. 371–395, 2017.

[33] D. G. Ko, F. Mai, Z. Shan, and D. Zhang, "Operational efficiency and patient-centered health care: A view from online physician reviews," *J. Operations Manage.*, vol. 65, no. 4, pp. 353–379, 2019.

[34] I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, and G. H. Gonzalez, "Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts," *J. Biomed. Informat.*, vol. 62, pp. 148–158, 2016.

[35] M. Emmert, F. Meier, A. K. Heider, C. Dürr, and U. Sander, "What do patients say about their physicians? An analysis of 3000 narrative comments posted on a German physician rating website," *Health Policy*, vol. 118, no. 1, pp. 66–73, 2014.
[36] Y. Jung, C. Hur, D. Jung, and M. Kim, "Identifying lkey hospital service quality factors in online health communities," *J. Med. Internet Res.*, vol. 17, no. 4, 2015, Art. no. e90.
[37] B. L. Ranard et al., "Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care," *Health Affairs*, vol. 35, no. 4, pp. 697–705, 2016.
[38] T. Bekhuis, M. Kreinacke, H. Spallek, M. Song, and J. A. O'Donnell, "Using natural language processing to enable in-depth analysis of clinical messages posted to an internet mailing list: A feasibility study," *J. Med. Internet Res.*, vol. 13, no. 4, 2011, Art. no. e98.
[39] A. T. Chen, "Exploring online support spaces: Using cluster analysis to examine breast cancer, diabetes and fibromyalgia support groups," *Patient Educ. Counseling*, vol. 87, no. 2, pp. 250–257, 2012.
[40] Y. Lu, P. Zhang, and S. Deng, "Exploring health-related topics in online health community using cluster analysis," in *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, 2013, pp. 802–811.
[41] A. Attard and N. S. Coulson, "A thematic analysis of patient communication in Parkinson's disease online support group discussion forums," *Comput. Hum. Behav.*, vol. 28, no. 2, pp. 500–506, 2012.
[42] K. Portier et al., "Understanding topics and sentiment in an online cancer survivor community," *J. Nat. Cancer Inst., Monographs*, vol. 2013, no. 47, pp. 195–198, 2013.
[43] Y. Zeng, X. Liu, Y. Wang, F. Shen, and H. Liu, "Recommending education materials for diabetic questions using information retrieval approaches," *J. Med. Internet Res.*, vol. 19, no. 10, 2017, Art. no. e342.
[44] H. Hao and K. Zhang, "The voice of Chinese health consumers: A text mining approach to web-based physician reviews," *J. Med. Internet Res.*, vol. 8, no. 5, 2016, Art. no. e108.
[45] X. Zhai, X. Feng, F. Guo, and Y. Huang, "The study on traditional Chinese medicine syndromes based on text mining method," in *Proc. IEEE 3rd Inf. Technol. Mechatron. Eng. Conf.*, 2017, pp. 777–780.
[46] T. Ruan et al., "An automatic approach for constructing a knowledge base of symptoms in Chinese," *J. Biomed. Semantics*, vol. 8, no. S1, 2017, Art. no. 33.
[47] G. Fang, L. Su, Q. Wang, and J. Wang, "Medical community expert classification based on potential semantic feature transfer learning," in *Proc. 29th Chin. Control Decis. Conf.*, 2017, pp. 5278–5281.
[48] D. Yang, Z. Yao, and R. Kraut, "Self-disclosure and channel difference in online health support group," in *Proc. 11th Int. Conf. Web Social Media*, 2017, pp. 704–707).
[49] X. Liu, Y. Zhou, and Z. Wang, "Deep neural network-based recognition of entities in Chinese online medical inquiry texts," *Future Gener. Comput. Syst.*, vol. 114, no. 1, pp. 581–604, 2020.
[50] D. Korencic, S. Ristov, and J. Snajder, "Document-based topic coherence measures for news media text," *Expert Syst. Appl.*, vol. 114, no. 11, pp. 357–373, 2018.
[51] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 101, no. 1, pp. 5228–5235, 2004.
[52] B. Leonid and C. O. Daniel, "Outlier detection for robust multi-dimensional scaling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2273–2279, Sep. 2019.



Xin Liu received the doctor's degree in management from the Business School, Central South University, Changsha, China, in 2022.

He is an Associate Professor of Intelligence Business Analytics & Data Science with the College of Economics and Management, Hengyang Normal University of China, Hengyang, China. His research interests include data analysis, feature engineering, mobile computing, service computing, deep learning, transfer learning, distributed computing, pervasive computing, and cloud computing. He is involved in many funded research projects as principal Investigator and technical members. He has published several research papers in reputed journals, including IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, *International Journal of Information Management*, *Information Processing & Management*, *Future Generation Computer Systems*, etc.

Yanju Zhou received the bachelor's degree in economics and the M.S. degree in management science from Central South University, Changsha, China, in 1996 and 2002, respectively, and the Ph.D. degree in management science from Beihang University, Beijing, China, in 2007.

She is a Professor and Doctoral Tutor. She is an expert in Decision Science and Supply Chain Management. She published about 30 articles in academic journals including *International Journal of Production Economics*, *Knowledge-Based Systems*, *Human and Ecological Risk Assessment: An International Journal*, *Journal of Intelligent & Fuzzy Systems* and so on.

Zongrun Wang received the Ph.D. degree in management from Central South University, Changsha, China, in 2004.

In 2014, he was the Vice Dean with the School of Business, Central South University. In 2010, a Professor and a Doctoral Supervisor, and in 2008–2009 a Visiting Scholar with California State University, Northridge. He has published some 40 articles in academic journals including *The Australian Economic Review*, *Economic Modelling*, *Journal of Applied Statistics*, *Physica A: Statistical Mechanics and its Applications*, *International Journal of Production*, among others, also three academic works. Much of his research is substantially supported by grants from National Natural Science Foundation of China. He is a peer Reviewer for *Journal of Financial Stability*, *Journal of Banking and Finance*, *International Journal of Production Economics*.

Ajay Kumar received the Ph.D. degree in business analytics from IIT Delhi, New Delhi, India, in 2017.

He is currently an Assistant Professor with EMLYON Business School, Ecully, France. He has been a Postdoctoral Fellow with Massachusetts Institute of Technology, Cambridge, MA, USA, and Harvard Business School, Boston, MA, USA. He has published several research papers in reputed journals, including IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, *Decision Support Systems*, *International Journal of Production Economics*, *Industrial Marketing Management*, *Technological Forecasting & Social Change*, *Annals of Operation Research*, *Journal of Business Research*, etc. His research and teaching interests include business analytics and data science.

Baidyanath Biswas received the Ph.D. degree in information systems from the Indian Institute of Management Lucknow, Lucknow, India, in 2019 with a specialization in cybersecurity and IT risk management.

He is currently an Assistant Professor of Business Analytics with the Trinity Business School, Dublin, Ireland. His research interests include IT risk management and business analytics. His research has appeared in *Decision Support Systems*, *Electronic Markets*, *Journal of Business Research*, and *Computers in Industrial Engineering*. He has nine years of rich industry experience as a Mainframe and DB2 database analyst at Infosys and IBM.

Dr. Biswas is associated with top peer-reviewed international conferences, namely, HICSS and ICIS. He currently serves as the Associate Editor of the *Electronic Markets* journal.