Check for
updates

Should doctors use or avoid medical terms? The influence of medical terms on service quality of E-health

Jilong Zhang[1] · **Jin Zhang**[1] · **Kanliang Wang**[1] · **Wei Yan**[2,3]

Abstract

With epidemics and pandemics like COVID-19, many offline healthcare services have been suspended and shifted to online, where patients and doctors typically communicate by typing texts. The limited communication poses a threat to the service quality of E-health, and also raises higher demand on the language skills of doctors, in which medical terms are a common concern. Traditional studies of offline healthcare mostly hold a negative attitude towards the use of medical terms by doctors. However, should we still advise doctors to avoid using medical terms in E-health? To answer this question, this paper conducts a study combining technical and empirical analyses based on real data. In this paper, a novel unsupervised text-mining method is proposed to automatically identify medical terms with crowd wisdom from large-scale doctor-patient communication texts. Then, a TREC-type experiment is carried out to validate the proposed method in terms of Precision, Recall, and F_1-measure, demonstrating that it can identify accurate and comprehensive medical terms. Based on the identified medical terms, an empirical analysis is conducted to verify the influence of medical terms used by doctors on the service quality of E-health. The analysis results show that for patients with low health literacy, the use of medical terms by doctors would decrease their service quality. However, for patients with high health literacy, the use of medical terms by doctors can significantly increase their service quality, revealing that doctors could improve their service quality in E-health by adjusting their medical term usage according to the health literacy of patients.

✉ Jin Zhang
   zhangjin@rmbs.ruc.edu.cn

Extended author information available on the last page of the article

1 Introduction

Owing to the widespread COVID-19, many traditional offline medical services have been suspended and shifted to online, which unexpectedly stimulates the development of E-health. For example, Ping An Good Doctor, a listed E-health company, announces that new registered users have increased by 10 times than usual, and the consultation times from new registered users have increased by 9 times.[1] In the first quarter of 2020, there are 390 million unique active users in the health channel of Alipay.[2] According to the report of China International Capital Corporation (CICC), the market size of E-health can reach 94 billion CNY in 2020.[3]

In E-health, service quality is a common concern for long-term development. Service quality measures how the delivered service level matches customer expectation [22, 40]. *"As an emerging service, the service standard of E-health remains unclear. The platform should optimize the procedure, train doctors, monitor and improve the service quality."* said by Wang Hang, the founder and CEO of Good Physician Online, a leading E-health platform with 610 thousand doctors. A speaker from WeDoctor, an E-health platform with 200 million users, expressed a similar concern that *"patients raise higher demands of service quality"*[4].

How to improve service quality in E-health? To answer this practical question, many academic efforts have been devoted to identifying factors that can affect service quality [31, 52, 57]. For this issue, existing studies mainly focus on the information of quantitative indexes, such as response speed, reply times [57], doctor titles and online experience [31], response rates and reply length [52]. However, few textual factors are identified from doctor-patient communication texts. In E-health, text-based electronic communication is an important and effective way [37]. The role of doctor-patient communication texts should be emphasized for the following two reasons: (1) medical services provided by doctors are mainly delivered through doctor-patient communication texts, which would influence the process of service [57]; (2) doctor-patient communication texts contain most of the medical information that patients seek for in E-health [28, 49], which would directly influence the service outcome [57]. The communication texts are received and processed by patients to retrieve medical knowledge and advice [49]. Regarding a patient as the receiver of textual information, it is a natural concern that whether the patient successfully receives enough information and sufficiently understands the information [49, 55] through doctor-patient communication texts, which indicates whether the information demand of patients is satisfied.

From the perspective of doctor-patient communication texts and the ability of patients to understand doctors, this paper focus on a specific factor of the medical

---

terms used by doctors. In the literature of doctor-patient communication, many studies concern the language used by doctors [2, 26]. Doctors can use either a simple or a professional expression to deliver similar information. For example, doctors may simply suggest patients to take "drugs", while they may also mean the same thing with specific professional drug names, such as "Motilium" or "Ibuprofen". In medical research, these professional expressions are called "medical terms" [2, 26, 39]. Medical terms are fundamental for patients to understand the nature of diseases/symptoms and the diagnostic process [39], but they are challenging for patients to understand [26, 39, 43]. Patients are believed low-literate in health knowledge [26, 39], which leads to a poor understanding of medical terms. Thus, most studies hold a conservative and cautious attitude towards using medical terms by doctors [19, 26, 39, 43], although it is impractical to totally avoid using medical terms in the process of medical services [2].

It is noteworthy that all those discussions about using medical terms are based on the traditional offline context, where patients are assumed to be low-literate. The online context of E-health provides patients with more access to medical information and knowledge. Patients are able to learn from the experience of other patients [1, 29, 56] and free knowledge from doctors [32, 60], which could improve the health literacy of some patients [45]. In addition, there exist numerous related documents on the Internet, in either a professional style or a popular style. Since patients and doctors communicate asynchronously in E-health, patients have enough time to query unfamiliar terms, which would ensure smooth communication for those patients who are willing to gain a better understanding of medical terms. Acknowledging the positive influence of E-health on patients' health literacy, it further obscures the influence of using medical terms on service quality in the online context. Thus, this paper aims to answer the following two research questions in E-health:

1. Does the use of medical terms still affect the service quality in E-health?
2. For patients with different levels of health literacy, should we encourage or discourage doctors to use medical terms in E-health?

The research questions above highly motivate this study to verify the influence of medical terms on service quality in E-health. In terms of identifying medical terms, existing studies [10] mainly rely on authority dictionaries, e.g. Unified Medical Language System (UMLS). However, they require a lot of manpower to maintain and update, and cannot respond quickly to the emergence of new terms. Moreover, they hardly cover all diseases in different language environments, and there are few wide-accepted authoritative dictionaries for some specific diseases, e.g. the heart disease of coronary atherosclerotic. Thus, in this study, we first propose an unsupervised medical term identification method with the crowd wisdom from Baidu Wikipedia based on text mining. With the identified medical terms, we can accurately and comprehensively figure out how many medical terms are used in a large corpus of communication texts between doctors and patients. Then, we further explore the influence of medical terms on service quality by conducting an empirical analysis

on a Chinese E-health website, Good Physician Online. The empirical results indicate that for patients with low health literacy, medical terms used by doctors would decrease their service quality. While for patients with high health literacy, medical terms used by doctors can significantly improve their service quality.

The contributions of this paper can be summarized as:

1. We propose a novel unsupervised text-mining method to identify medical terms from a large scale of doctor-patient communication texts with the crowd wisdom of Baidu Wikipedia.
2. Contrary to previous research findings, we find the positive influence of medical terms used by doctors on service quality for patients with high health literacy in E-health, while such a influence is significantly negative for patients with low health literacy.

The rest of this paper is structured as follows. Section 2 reviews related work. Section 3 develops the main hypotheses. Section 4 introduces the target E-health platform and the data used in this study. The proposed medical term identification method is presentd in Sect. 5. Section 6 elaborates the validation results for the method and the empirical results for the hypotheses. Finally, Sects. 7 and 8 respectively discuss the implication of this work and conclude the work with future research.

2 Literature review

2.1 Service quality in online healthcare service

In the literature of online healthcare services, many studies regard service quality as an antecedent to patients' behavior, such as adoption [7, 24, 47, 58], satisfaction and compliments [47, 52]. Adoption means whether patients choose online healthcare services [47] or certain doctors to consult [7, 58], while satisfaction and compliments are after-service evaluation regarding doctors and service processes [47, 52].

High service quality attracts more patients to seek for medical consultation, both online [7, 58] and offline [24]. Yang et al. revealed that the service quality indicators generated by patients positively influenced the search, evaluation and decision-making of patients, and encouraged patients to visit a doctor's homepage and consult the doctor [58]. In addition, Cao et al. found that the service quality and electronic word-of-mouth (eWOM) could affect patients' consultation decisions, which was moderated by the mortality and commonness of diseases [7]. Beyond online consultation, Li et al. found that high service quality facilitated patients to switch from online to offline medical services [24].

Service quality also influences patients' evaluation. Wu et al. examined the healthcare service quality at different stages (i.e. before-/during-/after-sales) and further explored how service quality influenced patients' psychological and compliment behavior [52]. Shim and Jo conducted an online survey and revealed that

service quality, information quality and system quality were promising to influence patients' perceived benefit of online health information sites, via mediators such as user satisfaction and intention to reuse [47]. In addition, Wu and Lu identified an inverted U-shape relationship of service prices and satisfaction, and found that service quality positively influenced patient satisfaction [53].

Although these studies validated the positive role of high-qualified services in patients' adoption, satisfaction and compliments, few studies aimed to provide practical instruments on how to improve service quality. Moreover, most studies only considered service quality in terms of response speed and response times [57]. Recently, a small number of studies attempt to answer the question of how to improve the service quality of E-health. For example, Wu and Deng investigated the knowledge coordination in online medical teams, which improved order quantities, patient satisfaction and service quality [51]. Liu et al. explored the influence of doctors' voice characteristics on patient satisfaction [30]. From a perspective of doctor-patient relationship, Zhang et al. emphasized the role of interaction unfairness on doctor-patient relationship quality [61], and Chen et al. [12] empirically validated that informational/emotional support from doctors improved patient satisfaction with service quality and attitude.

Although how to improve the quality of online healthcare services has received attention from some recent studies [12, 30, 51, 61], the semantic factors of online healthcare services still have not been properly explored. Since doctor-patient communication is the core of online healthcare services, whether patients can understand the doctor-patient communication could pose a great influence on service quality [19, 26]. Based on patients' understanding of doctor-patient communication, this study will explore the influence of the usage of medical terms by doctors on the service quality in E-health, aiming to enrich the research of service quality and provide practical suggestions for improving the service quality of E-health.

2.2 Patient understanding of medical terms

Since online healthcare service can be regarded as the information seeking process of patients [28], it is essential for patients to understand doctors' information through doctor-patient communication. In doctor-patient communication, many studies concern the language used by doctors [2, 26]. Specifically, from the perspective of patient understanding, medical terms used by doctors are often considered challenging to understand [26, 39, 43].

In existing literature, the use of medical terms is mainly discussed in offline situation, but rarely in online situation. In offline situation, many scholars conducted surveys to evaluate patients' understanding of medical terms. For example, O'Connell et al. found that patients' understanding of medical terms was fundamental to understanding the nature of diseases/symptoms and the diagnostic processes. However, patients were unable to understand the majority of medical terms, which might lead to the confusion, anxiety and breakdown of the doctor-patient relationship [39]. Pieterse et al. found lay patients were able to understand part of cancer-related terms, while few could understand all the test-related terms [42]. Schnitzler et al. revealed

that patients had difficulty in understanding medical terms in radiation oncology consultation, and felt embarrassed to admit that they failed to understand [43]. In pediatric surgical setting, Links et al. found that unexplained medical terms impeded the communication between doctors and patients' parents, which further discouraged the use of medical terms [26]. A recent study found the linguistic signals of messages would promote social support provision in online health community [11], and the authors measured the "readability" of messages based on FRE metric, which estimated the ease of reading with the number of words, syllables and sentences, rather than professional or lay terms.

All these above concerns about medical terms are raised on the precondition that patients have difficulty in understanding medical terms, which is a common assumption in offline situation. However, for patients' understanding of medical terms used by doctors, there remain research gaps. (1) Traditional studies generally conduct a qualitative analysis or a descriptive statistics analysis, with the absence of empirical studies on real data. (2) There are few studies that explore the role of using medical terms in the online context, where the traditional assumption that patients have difficulty in understanding medical terms is challenged [45]. Therefore, this study will empirically examine the influence of doctors' using medical terms on their service quality in E-health.

3 Hypothesis development

Service quality is a measure of how well the level of delivered services matches customer expectation [22]. For services, what customers receive may be entirely different from providers' intention [40]. Therefore, the interaction between providers and customers play a fundamental role in service quality [21].

In E-health, the interaction between service providers and receivers is in form of doctor-patient textual communication [12, 52, 57]. The communication texts are open to all visitors to obtain medical information they are interested in. For privacy concerns, all the patients are anonymous, which makes it impossible to track a single patient. Patients report their conditions and symptoms to doctors in textual description [57], and doctors provide information and support in textual responses [12, 52]. In this context, service quality of E-health is highly influenced with the doctor-patient communication processes [57]. Qualified medical consultation services depend on successful bi-directional communication with adequate and accurate understanding of each other. Nevertheless, a very challenging part in doctor-patient communication is medical terms [19, 43].

In offline situation, a majority of research efforts hold a negative attitude towards the use of medical terms, considering limited health literacy of patients [19, 26, 39, 43]. Health literacy is *the degree to which individuals can obtain, process, understand, and communicate about health-related information needed to make informed health decisions* [4]. Patients, generally with low health literacy, fail to understand or misunderstand doctors' explanations [19], prescription labels [33] and treatment plans [43].

However, the low-health literacy assumption of patients should be re-considered in the online situation, driven by three unique characteristics of E-health: (1) Active online engagement: Patients with high health literacy tend to actively engage in online health-related activities, such as health information seeking and use [62]. (2) Learning and sharing: Patients accumulate more knowledge from online activities. For example, active participation in online healthcare community helps members enhance their health literacy [36], since they could read a lot of popular medical contents written by doctors or experiences shared by other patients [1, 29, 32, 56, 60], which offers a good opportunity to get familiar with and understand related medical terms [10, 45]. (3) Asynchronous communication: E-health communication can span a long period of time. Patients do not need to respond immediately, and have ample opportunities to learn what doctors say between communication intervals. For example, patients can conduct timely search during the online communication, while they could only search for information after the offline communication in clinical offices [35]. Although E-health provides rich channels for patients to improve their health literacy, health literacy varies from patient to patient, which can be influenced by patients' education levels, the ability of using languages, past medical histories, residential locations, and poverty levels [34]. Such public information may further widen the knowledge gap among patients, which may cause different reactions to medical terms used by doctors.

Therefore, there is a great possibility that some patients could develop high health literacy in online situation, which requires re-thinking about the influence of doctors' using medical terms on service quality of E-health. Patients, with different health literacy, may have different expectations for the communication. For patients with low health literacy, they would prefer medical suggestions with simple language [19], despite the limited information transferred in simple languages. In contrast, patients with high health literacy would prefer in-depth discussion and advanced knowledge, and hope to get involved in decision-making [48, 62].

In conclusion, the influence of medical terms should be associated with patients' ability of understanding medical terms, i.e., the health literacy of patients. For patients with low health literacy, they have difficulty in understanding medical terms in doctor-patient communication texts. Therefore, the use of medical terms may disturb or impede patients' information acquisition, and subsequently decrease service quality. However, for patients with high health literacy, they are supposed to have smooth and accurate understanding of medical terms. Medical terms in their communication with doctors would deliver sufficient information without any ambiguity, and further improve service quality. Therefore, we have the following two hypotheses:

H1 In E-health, for doctors whose patients are of low health literacy, the use of medical terms by doctors would decrease their service quality.

H2 In E-health, for doctors whose patients are of high health literacy, the use of medical terms by doctors would increase their service quality.

4 Research context

To explore the influence of using medical terms on service quality in E-health, we collected real data on a typical Chinese E-health website, Good Physician Online, which is abbreviated as GoodPhy in this study. The platform of GoodPhy website is among the leading and most mature E-health websites in China. Until June 2020, the website has recorded basic information of more than 650,000 doctors from over 10,000 hospitals all over China. More than 232,000 doctors have been providing online health consultation services on this platform.

To get specific medical information and suggestions from E-health, typical patients firstly search for doctors experienced in certain diseases, then write down their self-description of symptoms online, and wait for the responses from doctors. Each online consultation is recorded in text form on the platform. We automatically crawled the consultation data between patients and doctors on the GoodPhy website. To eliminate the effect of different diseases, we selected 15 diseases, namely coronary heart disease, hepatitis B liver disease, hypertension, hyperlipidemia, insomnia, headache, gastritis, leukemia, cirrhosis, cerebral infarction, nephritis, diabetes, arrhythmia, irregular menstruation, and pancreatitis, which are commonly discussed in existing literature of E-health [57]. For each disease, the GoodPhy platform provides a recommended list of doctors, from which we chose the top 300 doctors. By removing duplicates, we totally got 3594 unique doctors. Due to privacy protection, some doctors chose to make their consultation logs private. Thus, we only kept doctors who chose to publish their consultation logs. For each remained doctor, we collected all the consultation data between the doctor and his/her patients before February 2018, which is summarized in Table 1. In total, 117,196 records of communication were crawled (66.1 records per doctor on average). To avoid the influence of insufficient data, we excluded two diseases hyperlipidemia and arrhythmia that ranked last in the list of cases per doctor. All the retained diseases are denoted as a set *Disease*.

5 Medical terms identification

Healthcare-related texts involve abundant professional medical terms, including words and phrases. To extract terms, traditional studies mainly depend on dictionaries generated by experts. However, such dictionaries require long-time subjective efforts for compilation and revision, which can hardly keep pace with the rapid development of medicine. In addition, there are few wide-accepted authoritative dictionaries for some specific diseases, such as the heart disease of coronary atherosclerotic. Therefore, considering the large amount of communication texts between doctors and patients for various diseases, an effective way is to automatically extract disease-specific medical terms with text mining technique.

In this study, we propose an unsupervised method to identify medical terms form the communication texts between doctors and patients, which is presented

Table 1 Summary of the GoodPhy data used in this study

Disease	# Doctors with consultation logs	# Consultation cases	# Cases per doctor
Hepatitis B liver disease	118	17,869	151.43
Hypertension	127	17,168	135.18
Diabetes	131	13,643	104.15
Irregular menstruation	200	19,510	97.55
Coronary heart disease	101	9229	91.38
Gastritis	112	9022	80.55
Nephritis	96	6127	63.82
Insomnia	153	7873	51.46
Cirrhosis	111	4707	42.41
Headache	133	4031	30.31
Cerebral infarction	127	3631	28.59
Leukemia	67	1570	23.43
Pancreatitis	115	1606	13.97
Hyperlipidemia	99	802	8.10
Arrhythmia	83	408	4.92
Total	1773	117,196	66.10

Fig. 1 The method framework
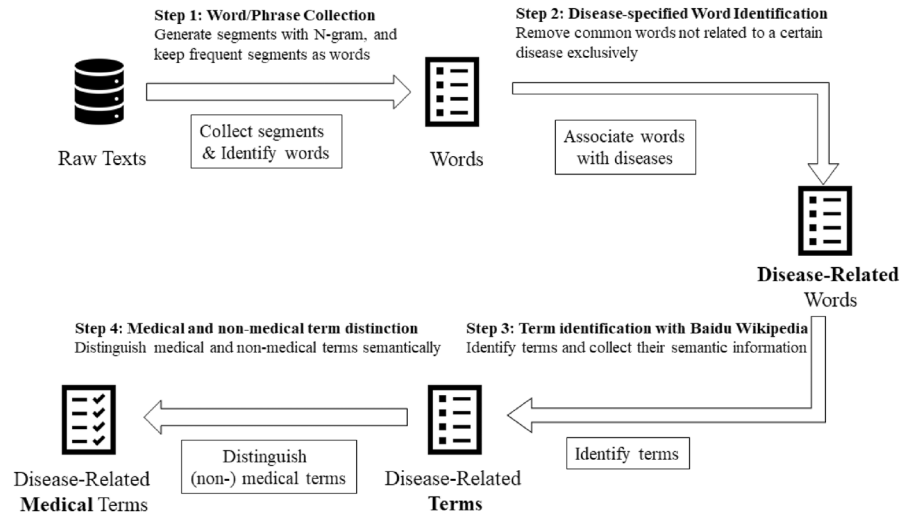
in Fig. 1. Medical terms are first and foremost related to a specified disease. In the meanwhile, they should be recognized by the public. Therefore, the process of medical term identification is designed as follows. Briefly, we firstly extract N-gram segments from original texts and collect words and phrases by identifying frequent segments. Secondly, we further remove common words/phrases

Fig. 2 An example for N-gram

Raw text	在胃镜检查中也没找到出血点

2-gram

在胃/胃镜/镜检/检查/查中/中也/
也没/没找/找到/到出/出血/血点

3-gram

在胃镜/胃镜检/镜检查/检查中/查中也/
中也没/也没找/没找到/找到出/到出血/出血点

4-gram

在胃镜检/胃镜检查/镜检查中/
检查中也/查中也没/中也没找
/也没找到/没找到出/找到出血/到出血点

Manual Split

在/胃镜检查/中/也/没/找到/出血点
didn't find petechia in gastroscopy neither

that are not disease-specified. Thirdly, we apply Baidu Wikipedia to identify terms from words/phrases and collect their semantic corpus. Only the terms recognized and included in Baidu Wikipedia are preserved. Next, based on Baidu Wikipedia corpus for each term, we build the Doc2vec model [20] to extract the knowledge-based semantic expression of each term. Finally, based on the semantic expressions of terms, we use K-means to cluster all the terms to identify the cluster containing medical terms. The detailed process of the proposed method is described as follows.

5.1 Word/phrase collection

In the first step, initial words and phrases are extracted from online communication texts between doctors and patients. This study is conducted on a Chinese E-health platform where the communication texts are in Chinese. In Chinese, there is no explicit separator of words/phrases in sentences. Therefore, the first challenge is Chinese word segmentation. Considering all potential situations, we adopt the N-gram segmentation method on the communication texts. For each text in doctor-patient communication concerning $disease_i \in Disease$, the text is split for each N-gram length (2–7 tokens). Figure 2 shows an example for N-gram processing.

The results of N-gram include words, phrases, and some meaningless combination of Chinese characters. In Fig. 2, we translate the N-gram results into English, where each color represents the Chinese/English expressions with the same meaning. It can be found that some character combinations generated by the N-gram segmentation method are meaningless. To filter those segments, we only preserve segments which appear more than 5 times. For $disease_i \in Disease$, all the retained segments compose a set of words, denoted as $Words_{i_{all}}$.

Fig. 3 An example for word polysemy in Baidu Wikipedia

5.2 Disease-specified word identification

In the set of words $Words_{i_{all}}$, there are still some meaningful but common words, such as "hospital" and "doctor", which could be used for most diseases and cannot be considered as disease-specified words. Therefore, we need to filter out those words in $Words_{i_{all}}$ through an excluding process. In general, words/phrases truly concerning a specific disease $disease_i$ should not appear in the communication texts about other diseases. Therefore, the disease-specified words of $disease_i$, denoted as $Words_{i_{cand}}$, can be defined as Eq. 1. Based on the words collected by the N-gram method, the set of disease-specific words is then built up.

$$Words_{i_{cand}} = Words_{i_{all}} \setminus \bigcap_{j \neq i} Words_{j_{all}} \tag{1}$$

5.3 Term identification with Baidu Wikipedia

In the next step, each word in the set of disease-specific words for $disease_i$ is searched in Baidu Wikipedia, and only words existing in Baidu Wikipedia are deemed as disease-specific terms. For each word $word_i \in Words_{i_{cand}}$, the returned Baidu Wikipedia contents (if exist) are downloaded as the characteristic document of the word. If more than one entries are retrieved for a word(the number of entries is denoted as $N = N_0$.), it means that there exist polysemous terms. In this case, all the contents of the word are crawled separately, and multiple terms are created correspondingly, denoted by $term_{ij}$ ($j = 1, ..., N_0$). For example, as shown in Fig. 3, there exist two entries named Dopamine in Baidu Wikipedia. The first entry of Dopamine means the organic chemical in the brain as a neurotransmitter. The second entry of Dopamine refers to the romantic film directed by Mark Decena. Therefore, we could identify two terms for the word Dopamine via Baidu

Wikipedia, and further generate two separate characteristic documents for these two terms.

5.4 Medical and non-medical term distinction

For *disease_i*, all the identified terms through Baidu Wikipedia form the set of terms *Terms_i*, which are recognized by the public. In *Terms_i*, there are still some elements which are disease-specified terms but not medical terms. For example, "furious" is a disease-specified term for the coronary heart disease and is also recognized as a term by the public, but it does not describe the nature of diseases or the diagnostic process, and should not be simply treated as a medical term. Thus, in this step, medical terms and non-medical terms are further distinguished through semantic analysis where we apply the deep-learning based Doc2vec model [20]. In the implementation of the Doc2vec model, the Chinese word segmentation is conducted with a widely-used Python package, Jieba [54].

We firstly train the Doc2vec model with the characteristic documents of terms. Considering that such complicated model requires a large scale of corpus, we use the characteristic documents of terms for all diseases to train the model. After the training process, the embedding vector of each term can be learned, which reflects the semantic expression of its characteristic document. In general, medical terms should be mapped close to each other in the vector space. Based on the characteristic vectors, the K-means clustering is implemented to cluster all the terms. In K-means, the number of clusters *K* should be decided in advance. In this study, pre-models are built with cluster numbers varying from 2 to 30, and the appropriate cluster number is determined according to the metric of semipartial *R*-squared (SPR) [46]. The cluster containing medical terms is then picked up manually. In the cluster, we then divide the identified terms to match different diseases.

The above four steps constitute the whole method, based on which medical terms can be automatically identified from a large number of doctor-patient communication texts for a given disease.

5.5 An illustrative example

To explain the process of medical term identification clearly, we take the disease of leukemia as an illustration. From the recommendation list of doctors experienced in leukemia on the GoodPhy website, we collected information of the top 300 doctors, and only 67 doctors chose to disclose their consultation logs about leukemia. 1570 consultation cases were recorded in total (23.43 consultations per doctor on average). Based on communication texts in these consultation cases, 1,281,454 segments were collected by the N-gram method, 49,112 of which appeared more than 5 times. After removing common words appearing in the consultation logs of other diseases, 8786 disease-specified words were preserved. Through analyzing the corpus of Baidu Wikipedia, only 394 words were identified as terms recognized by the public, which included 948 entries in total. By combining characteristic documents of terms with other diseases, we built a Doc2vec model with 16608 terms, and learn
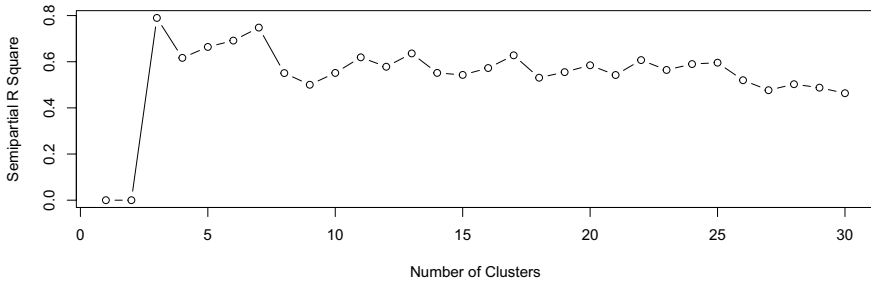
Fig. 4 Semipartial *R*-squared for pre-models on leukemia

the characteristic vector of each term. To find out the most appropriate cluster number, pre-models were built with cluster numbers varying from 2 to 30. The values of semipartial *R*-squared (SPR) for all the pre-models are displayed as Fig. 4, according to which the cluster number was decided as 3. The numbers of terms in these clusters were 6878, 3898, and 5832 respectively. These 3 clusters showed significant difference, among which the second cluster with 3898 entries was labeled as the one containing medical terms with manual judgment, and the other two clusters contained non-medical terms.

To clearly illustrate the identified medical terms of leukemia in the second cluster, representative terms in the cluster are exhibited with their Chinese expressions and English meanings in Fig. 5. According to Fig. 5, all the identified medical terms can be broadly divided into 5 categories, including diseases, symptoms, drugs, examinations and treatments. We can find that these medical terms are very specific to the disease of leukemia. Moreover, they are all professional expressions from different aspects of this disease, which are well recognized by the public.

6 Experiment and analysis

This section mainly focuses on two parts of experimental analyses. The first part validates the effectiveness of medical term identification with the proposed method. Based on the identified medical terms, the second part is to verify the hypotheses developed in Sect. 3 with an empirical analysis. The details of these two parts are elaborated in the following subsections

6.1 Medical term identification results

Since there is no standard evaluation dataset or ground truth data for medical terms of a given disease, we adopted the TREC-type evaluation framework to generate the ground truth labels of medical terms in the performance evaluation of the proposed method. TREC framework is one of the most classical evaluation methods, which is widely used in many tasks like information retrieval [50], comparable entity identification [3], etc. In the experiment, all the terms grouped into

Category	Chinese	English
Disease	小儿白血病	Pediatric leukemia
	慢性白血病	Chronic leukemia
	慢淋	Chronic lymphocytic leukemia
	急性髓系白血病	Acute myeloid leukemia
	cml	Chronic myelocytic leukemia
	浆细胞白血病	Plasma cell leukemia
	败血症	Septicemia
Symptom	白细胞增多	Leukocytosis
	嗜酸粒细胞	Basophilic granulocyte basophil
	肺部真菌感染	Pulmonary fungal infections
	骨髓抑制	Myelosuppression
	T细胞	T-lymphocyte
	血融	Blood melting
Drug	百炎净	Drug name of Compound Sulfamethoxazole Tablets
	枢丹	Drug name of Ondansetron hydrochloride injection
	新赛斯平	Durg name of Cyclosporine oral solution
	格列卫	Drug name of Imatinib mesylate tablets
	伊曲康唑	Drug name of Itraconazole
	长春新碱	Drug name of Vincristine
	万古霉素	Vancomycin
	美罗培南	Meropenem
Examination	骨髓穿刺	bone marrow puncture
	骨髓涂片	Bone marrow smear
Treatment	造血干细胞移植	Hematopoietic stem cell transplantation
	骨髓配型	Bone marrow transplantation
	靶向治疗	Targeted therapy

Fig. 5 Representative terms in the cluster of medical terms

medical or non-medical clusters, were regarded as the whole population. Terms in medical clusters could be seen as positive ones, while others were regarded as negative. To get the ground truth, three medical students were invited to identify whether a term in the population was a medical or non-medical term. Only the terms labeled as medical by at least two evaluators were considered as qualified medical terms. The label results generated by these three professional evaluators were regarded as the ground truth, based on which we could evaluate the performance of the proposed medical term identification method with typical metrics of Precision, Recall, and F_1-measure [8, 9].

Precision is the fraction of qualified medical terms in the whole population, and could indicate the accuracy of the proposed method in detecting qualified medical terms. Given a disease $disease_i$, the set of medical terms identified by the proposed method are denoted as $MedicalTerms_i$, and $LabeledMedicalTerms_i$ is the set of qualified medical terms labeled by evaluators for $disease_i$. The Precision value $Precision_i$ for $disease_i$ is formulated as:

Table 2 Medical term identification results with different Doc2vec vector dimensions

Vector dimension	#Terms	#True terms	#Positive terms	#True positive terms	Precision	Recall	F1
3	8814	3605	**3656**	**3090**	**0.845**	**0.857**	**0.851**
5	8814	3605	3659	3072	0.840	0.852	0.846
10	8814	3605	3483	2963	0.851	0.822	0.836
100	8814	3605	3379	2848	0.843	0.790	0.816

The line in bold highlights the best performance with the vector dimension of 3

$$Precision_i = \frac{\left|LabeledMedicalTerms_i \cap MedicalTerms_i\right|}{\left|MedicalTerms_i\right|} \tag{2}$$

where $|*|$ represent the scale of a set $*$.

Recall is the ratio of qualified medical terms identified by the proposed method over all qualified medical terms. It indicates the comprehensiveness of the identification results for the proposed method. For $disease_i$, the Recall value $Recall_i$ is formulated as:

$$Recall_i = \frac{\left|LabeledMedicalTerms_i \cap MedicalTerms_i\right|}{\left|LabeledMedicalTerms_i\right|} \tag{3}$$

F_1-measure which is the harmonic mean of Precision and Recall, could comprehensively evaluate the performance of the proposed method. It is formulated as:

$$F_{1i} = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{4}$$

The Doc2vec model was applied with the Gensim package in Python. We tried different values of vector dimensions, and their corresponding performance of medical term identification in terms of Precision, Recall, and F_1 metrics are shown in the Table 2. In Table 2, "True Medical Terms" is the number of qualified medical terms labeled by the evaluators, while "Positive Medical Terms" is the number of medical terms identified by the proposed method. "True Positive Medical Terms" is the number of identified medical terms which are also labeled by the evaluators as qualified medical terms. From the results of Table 2, when setting the vector dimension to 3, the model achieves best performance. The precision with the optimal vector dimension reaches 84.5%, and in the meanwhile, the proposed method can find 85.7% medical terms from all the true medical terms labeled by the evaluators, which demonstrates that the medical terms identified by the proposed method could be regarded as reliable for further empirical analysis [15]. Furthermore, Table 3 illustrates detailed results for all the diseases in the set of *Disease*.

We further evaluated the performance of the proposed method against benchmark methods. Firstly, we introduced a naive benchmark by simplifying the proposed method. The naive benchmark suggested the results obtained only through the first three steps of the proposed method, i.e., all the terms identified with

Table 3 The performance of medical term identification with the Doc2vec dimension of 3

disease	# Terms	# True medical terms	# Positive medical Terms	# True positive medical terms	Precision	Recall	F_1
Coronary heart disease	686	204	203	162	0.798	0.794	0.796
Diabetes	697	285	307	237	0.772	0.832	0.801
Cerebral infarction	251	124	115	106	0.922	0.855	0.887
Leukemia	393	218	178	165	0.927	0.757	0.833
Cirrhosis	175	76	79	65	0.823	0.855	0.839
Nephritis	605	379	367	331	0.902	0.873	0.887
Pancreatitis	115	61	55	52	0.945	0.852	0.897
Irregular menstruation	2202	805	953	778	0.816	0.966	0.885
Hypertension	429	209	173	157	0.908	0.751	0.822
Insomnia	494	174	173	139	0.803	0.799	0.801
Gastritis	785	471	432	403	0.933	0.856	0.893
Headache	193	106	110	95	0.864	0.896	0.880
Hepatitis B liver disease	1789	493	511	400	0.783	0.811	0.797
Total	8814	3605	3656	3090	0.845	0.857	0.851

Baidu Wikipedia were suggested as medical terms. Secondly, to validate the advantage of Doc2vec, we replaced Doc2vec with the LDA model by using the topic distributions as characteristic vectors. For LDA, we set the number of topics as 3, 10, 50, and 100 separately. Thirdly, to validate the advantage of clustering, we applied two alternative methods, namely one-class SVM (OCSVM) [44] and Isolation Forest [27]. One-class SVM is a novelty detection model and Isolation Forest is a outlier detection model. In these two models, both the identified novel terms and the outliers were suggested as medical terms respectively. The comparative results between the proposed method and benchmarks are shown in Table 4.

From Table 4, the proposed method achieves the best performance in terms of F_1-measure. For LDA models, the number of topics has a weak influence on the identification performance. Even the Naive method achieves a Precision score of 0.409, which reflects the validity of the previous three steps of the proposed method. For OCSVM and Isolation Forest, the -out/-in methods mean that the identified outlier/novelty points were suggested as medical terms. It can be found that the -out/-in methods are worse than the Naive method, which suggests that novelty/outlier detection methods are inferior to clustering in this task.

6.2 Main variable and summary statistics

Since we can accurately and comprehensively identify and extract medical terms from doctor-patient communication texts, we can further explore the influence

Table 4 The performance of proposed method against benchmarks

Vector dimension	#Terms	#True terms	#Positive terms	#True positive terms	Precision	Recall	F_1
Proposed	8814	3605	3656	3090	0.845	0.857	0.851
LDA N=10	8814	3605	3755	3106	0.827	0.862	0.844
LDA N=50	8814	3605	3581	3015	0.842	0.836	0.839
LDA N=100	8814	3605	3670	3045	0.830	0.845	0.837
LDA N=3	8814	3605	4075	3213	0.788	0.891	0.837
Naive	8814	3605	8814	3605	0.409	1.000	0.581
OCSVM-out	8814	3605	5007	2456	0.491	0.681	0.570
IsolationForest-out	8814	3605	2110	1317	0.624	0.365	0.461
IsolationForest-in	8814	3605	7201	2458	0.341	0.682	0.455
OCSVM-in	8814	3605	4717	1394	0.296	0.387	0.335

of medical terms used by doctors on their service quality in E-health. Removing doctors with no communication texts, we finally got 1467 samples. Each sample was associated with a doctor, a patient, and a specified disease. For a certain disease, all the doctors with communication texts related to this disease were selected as the set of samples. Table 5 and Table 6 present the variable description and summary statistics of variables.

There exist three voluntary ways for patients to provide positive feedback for the consultation services on the website of GoodPhy, namely voting, sending a thank-you letter and giving a digital gift. The number of votes, thank-you letters and digital gifts are regarded as "the information generated by patients reflecting doctors' service quality" [24, 58]. Therefore, we used these three numbers as dependent variables (DV) separately, denoted as N_Vote (the number of votes), N_Thank (the number of thank-you letters) and N_Gift (the number of digital gifts). The results with different DVs are also believed to provide robust conclusions jointly.

In this study, we conducted the empirical analysis from the doctor perspective for two reasons. Firstly, it is impractical to obtain information from each individual patient for privacy concerns. Moreover, the measurements of service quality are aggregated at the doctor level.

The independent variables, *Doc_Term_Freq* and *Pat_Term_Freq*, measure the frequency of using medical terms by a doctors and the patients of the doctor. Such measurements could reflect the levels of healthcare-related expertise [16] and health literacy [10]. We scaled the variables of both *Doc_Term_Freq* and *Pat_Term_Freq* through z-score. To avoid the influence of outliers, values larger than 3 were set to 3, and those less than -3 were set to -3 [59]. We also discussed the way of handling outliers in robustness check.

The control variables are consistent with an existing study of E-health [57]. We adopted the quantitative indexes in the doctor-patient communication , including response speed, reply frequency, titles of doctors and the risk of diseases [57].

Table 5 Variable description

Dimension	Variable	Definition
Dependent variables	N_Vote	The number of votes received by a doctor
	N_Thank	The number of thank-you letters received by a doctor
	N_Gift	The number of digital gifts received by a doctor
Independent variables	Doc_Term_Freq	Supposing that a doctor provided services to P patients, r_p denotes the number of times the doctor replied to $patient_p$, and in each $reply_{pi}$ by the doctor to $patient_p$, there were x_{pi} medical terms then $Doc\_Term\_Freq = \sum_{p=1}^{P}[\sum_{i=1}^{r_p} x_{pi}/r_p]/P$.
	Pat_Term_Freq	Supposing that a doctor provided services to P patients m_p denotes the number of messages sent by $patient_p$ and in each $message_{pj}$ by $patient_p$, there were y_{pj} medical terms then $Pat\_Term\_Freq = \sum_{p=1}^{P}[\sum_{j=1}^{m_p} y_{pj}/m_p]/P$
Control variables	Reply_Frequency	Supposing that a doctor provided services to P patients r_p denotes the number of times the doctor replied to $patient_p$ $Reply\_Frequency = \sum_{p=1}^{P} r_p/P$
	Reply_Speed	Supposing that a doctor provided services to P patients r_p denotes the number of times the doctor replied to $patient_p$ for each $reply_{pi}$ by the doctor to $patient_p$, it was sent at $answertime_{pi}$ and the closest message from $patient_p$ before $reply_{pi}$ was sent at $askingtime_{pi}$ $Reply\_Speed = \sum_{p=1}^{P}[\sum_{i=1}^{r_p} (answertime_{pi} - askingtime_{pi})/r_p]/P$
	Risk	If a disease is of high risk, Risk = 1, otherwise Risk = 0
	Title	The sum of a doctor's ranking and academic title. Ranking: director physician=4, associate director physician=3, chief physician=2, and physician=1 Academic title: namely professor=3, associate professor=2, and lecturer=1

Table 6 Summary statistics of variables

Variable	Min	Max	Mean	Std
N_Vote	2	1484	116.400	135.620
N_Thank	1	1068	64.700	92.121
N_Gift	1	5061	275.800	463.522
Doc_Term_Freq	0	1.1392	0.060	0.102
Pat_Term_Freq	0	1.14598	0.089	0.123
Reply_Frequency	1	156.857	4.333	8.074
Response_Speed	0	65.0667	2.609	4.744
Risk	0	1	0.472	0.499
Title	1	7	5.180	1.650

6.3 Empirical results

Since the dependent variables were measured as count data, traditional OLS model may lead to biased and inconsistent results [6, 41]. Instead, Poisson regression [41] and Negative Binomial regression [13] are models commonly used to analyze count data. Poisson regression requires the assumption of mean-variance equality [41], and data with great dispersion would lead to inefficient estimates [13]. To avoid the influence of overdispersion, Negative Binomial regression model introduces an idiosyncratic error term that captures unobserved characteristics [13]. According to Table 6, the variance of variables is larger than their mean values. We further conducted dispersion tests with the AER package [18, 41], which rejected the mean-variance equality assumption. Therefore, we used Negative Binomial regression (NB) to test our hypotheses. We also considered OLS regression and Poisson regression as robust check in Sect. 6.4.

Table 7 presents the estimation results of Negative Binomial regression. The VIF (Variance inflation factor) test indicates that there is no multi-collinearity problem in all the models (VIF< 1.6).

According to Table 7, s_Doc_Term_Freq has a significant negative effect on the number of thank-you letters (β = -0.095, $p < 0.05$), but the effect on the numbers of votes and gifts is not significant. However, the coefficient of the interaction term ($s\_Doc\_Term\_Freq \times s\_Pat\_Term\_Freq$) is significant for all the three indicators of service quality (β =0.086, $p < 0.001$ for N_Vote; β = 0.137, $p < 0.001$ for N_Thank; and β =0.107, $p < 0.01$ for N_Gift). The results indicate that as the health literacy of patients($s\_Pat\_Term\_Freq$) increases, the usage of medical terms by doctors ($s\_Doc\_Term\_Freq$) would significantly improve their service quality; on the contrary, as the health literacy of patients($s\_Pat\_Term\_Freq$) decreases, the negative influence of doctors' using of medical terms would be enhanced. Therefore, hypotheses H1 and H2 are supported.

For control variables, we validated the positive influence of the response speed and titles of doctors. Across three indicators of service quality, the response speed of doctors significantly influences service quality (β =0.026, $p < 0.001$ for N_Vote; β = 0.026, $p < 0.001$ for N_Thank; and β =0.029, $p < 0.001$ for N_Gift). The titles

Table 7 Parameter estimation: negative binomial regression

| | Dependent variables | | | | | | | | |
| | N_Vote | | | N_Thank | | | N_Gift | | |
	β	std	Sig	β	std	Sig	β	std	Sig
Constant	4.123	0.082	***	3.262	0.090	***	4.598	0.106	***
s_Doc_Term_Freq	− 0.034	0.034		− 0.095	0.037	*	− 0.004	0.043	
s_Pat_Term_Freq	0.121	0.030	***	0.151	0.033	***	0.199	0.039	***
Reply_Frequency	− 0.001	0.003		− 0.006	0.003	*	0.032	0.004	***
Response_Speed	0.026	0.005	***	0.026	0.005	***	0.029	0.006	***
Risk	− 0.059	0.046		0.028	0.051		− 0.081	0.060	
Title	0.103	0.014	***	0.146	0.015	***	0.128	0.018	***
s_Doc_Term_Freq× s_Pat_Term_Freq	0.086	0.026	***	0.137	0.028	***	0.107	0.034	**

Significance: *** $p < 0.001$; ** $p < 0.01$; *$p < 0.05$

s_Doc_Term_Freq: scale(Doc_Term_Freq)

s_Pat_Term_Freq: scale(Pat_Term_Freq)

of doctors also have positive affect on service quality ($\beta = 0.103$, p < 0.001 for N_Vote; $\beta = 0.146$, p < 0.001 for N_Thank; and $\beta = 0.128$, p < 0.001 for N_Gift). These results are consistent with those in the previous studies discussing response speed [57] and titles [31]

To sum up, the empirical results support hypotheses H1 and H2. For doctors whose patients are of low health literacy, the use of medical terms can decrease their service quality, and for doctors whose patients are of high health literacy, the use of medical terms by doctors can increase their service quality. For H1, it is consistent with the conclusion derived from the offline situation where patients with low health literacy prefer plain language than medical terms [19]. For H2, it is in line with the effective role of medical terms in delivering information [2] for capable patients. To further validate the difference effects on patients with high/low health literacy, we conducted the grouped regression in robustness check, as well as the robust check for the outliers and estimation methods in Sect. 6.4.

6.4 Robustness check

It is noteworthy that in the empirical results elaborated above, we used three indexes to measure service quality, namely the number of votes, thank-you letters and digital gifts. Across different measurements, we got consistent results in Table 7, which reflect that the results are robust to the measurement of service quality.

In this subsection, we conducted additional experiments to check the robustness of our results from four aspects: (1) we used the grouped regression to validate the difference between patients with high/low health literacy; (2) we tried alternative ways to deal with outliers; (3) we used alternative estimation models,

Table 8 Parameter estimation with negative binomial regression for doctors whose patients are of high health literacy

	Dependent variables								
	N_Vote			N_Thank			N_Gift		
	β	Std	Sig	β	Std	Sig	β	Std	Sig
Constant	4.317	0.112	***	3.447	0.126	***	4.531	0.141	***
s_Doc_Term_Freq	0.087	0.032	**	0.106	0.036	**	0.172	0.040	***
Reply_Frequency	0.038	0.007	***	0.031	0.008	***	0.110	0.009	***
Response_Speed	0.022	0.007	**	0.030	0.008	***	0.036	0.009	***
Risk	− 0.188	0.063	**	− 0.069	0.071		− 0.263	0.079	***
Title	0.057	0.018	**	0.101	0.021	***	0.092	0.023	***

Significance: *** $p < 0.001$; ** $p < 0.01$; *$p < 0.05$

Table 9 Parameter estimation with negative binomial regression for doctors whose patients are of low health literacy

	Dependent variables								
	N_Vote			N_Thank			N_Gift		
	β	Std	Sig	β	std	Sig	β	std	Sig
Constant	3.648	0.121	***	2.831	0.132	***	4.315	0.158	***
s_Doc_Term_Freq	− 0.229	0.071	**	− 0.237	0.077	**	− 0.120	0.093	
Reply_Frequency	− 0.005	0.003	.	− 0.011	0.004	**	0.012	0.004	**
Response_Speed	0.024	0.006	***	0.020	0.007	**	0.014	0.008	.
Risk	− 0.016	0.068		0.034	0.074		− 0.101	0.089	
Title	0.168	0.020	***	0.211	0.022	***	0.188	0.027	***

Significance: *** $p < 0.001$; ** $p < 0.01$; *$p < 0.05$; $p < 0.1$

such as OLS and Poisson regression models instead of Negative Binomial regression model; and (4) we conducted analysis in the case level, which combined the doctor fixed effect and Propensity Score Matching (PSM) to control the unobservable characteristics of doctors and overcome the selection bias.

Firstly, we divided doctors into two groups according to the health literacy of their patients. For doctors whose patients were of high/low health literacy (over/below the median), they were assigned as the high/low group. We re-ran Negative Binomial model on these two groups.

According to Tables 8 and 9, in the high group, doctors' using medical terms significantly improves their service quality measured by N_Vote, N_Thank or N_Gift. On the other hand, in the low group, doctors' using medical terms significantly reduces the number of votes and thank-you letters. The influence of Doc_Term_Freq on N_Gift is not significant in the low group, but it is significantly

Table 10 Parameter estimation with Negative Binomial regression by keeping outliers

	Dependent variables								
	N_Vote			N_Thank			N_Gift		
	β	Std	Sig	β	Std	Sig	β	Std	Sig
Constant	4.152	0.081	***	3.297	0.089	***	4.633	0.105	***
s_Doc_Term_Freq	− 0.019	0.030		− 0.072	0.033	*	0.006	0.038	
s_Pat_Term_Freq	0.122	0.026	***	0.153	0.029	***	0.200	0.034	***
Reply_Frequency	− 0.001	0.003		− 0.007	0.003	*	0.032	0.004	***
Response_Speed	0.026	0.005	***	0.026	0.005	***	0.028	0.006	***
Risk	− 0.071	0.046		0.015	0.051		− 0.094	0.060	
Title	0.103	0.014	***	0.145	0.015	***	0.127	0.018	***
s_Doc_Term_Freq×s_Pat_Term_Freq	0.026	0.014	. ($p = 0.069$)	0.061	0.016	***	0.039	0.019	*

Significance: *** $p < 0.001$; ** $p < 0.01$; *$p < 0.05$; $p < 0.1$

Table 11 Parameter estimation with Negative Binomial regression by removing outliers

	Dependent Variables								
	N_Vote			N_Thank			N_Gift		
	β	Std	Sig	β	Std	Sig	β	Std	Sig
Constant	4.093	0.083	***	3.230	0.091	***	4.592	0.107	***
s_Doc_Term_Freq	− 0.042	0.037		− 0.110	0.041	**	0.004	0.048	
s_Pat_Term_Freq	0.131	0.034	***	0.148	0.037	***	0.185	0.044	***
Reply_Frequency	− 0.001	0.003		− 0.007	0.003	*	0.030	0.004	***
Response_Speed	0.025	0.005	***	0.024	0.005	***	0.027	0.006	***
Risk	− 0.045	0.047		0.043	0.052		− 0.072	0.061	
Title	0.106	0.014	***	0.150	0.015	***	0.130	0.018	***
s_Doc_Term_Freq×s_Pat_Term_Freq	0.119	0.035	***	0.155	0.038	***	0.117	0.045	**

Significance: *** $p < 0.001$; ** $p < 0.01$; *$p < 0.05$

positive in the high group. In summary, the results of the grouped regression also support the hypotheses of H1 and H2.

Secondly, we examined that whether the results were robust to the method of handling outliers. In the main analysis, we avoided the potential influence of outliers in Doc_Term_Freq and Pat_Term_Freq by setting all the standardized scores larger than 3 to 3 and those lower than -3 to -3. In the robustness check, we tried two alternative ways to handle outliers: (1) we did not handle outliers and just kept them as the standardized scores. (2) we directly removed all the outliers from the dataset. The results are shown in Tables 10 and 11, which are consistent with the results in Table 7.

Table 12 Parameter estimation with OLS regression (scaled DV)

	Dependent variables								
	s_N_Vote			s_N_Thank			s_N_Gift		
	β	Std	Sig	β	Std	Sig	β	Std	Sig
Constant	− 0.476	0.092	***	− 0.574	0.091	***	− 0.472	0.091	***
s_Doc_Term_Freq	− 0.028	0.038		− 0.074	0.037	*	− 0.004	0.037	
s_Pat_Term_Freq	0.094	0.034	**	0.105	0.033	**	0.157	0.033	***
Reply_Frequency	0.000	0.003		− 0.002	0.003		0.015	0.003	***
Response_Speed	0.022	0.005	***	0.021	0.005	***	0.015	0.005	**
Risk	− 0.039	0.052		0.035	0.051		− 0.009	0.051	
Title	0.077	0.016	***	0.089	0.015	***	0.060	0.015	***
s_Doc_Term_Freq× s_Pat_Term_Freq	0.100	0.029	***	0.149	0.029	***	0.107	0.029	***

Significance: *** $p < 0.001$; ** $p < 0.01$; *$p < 0.05$

Table 13 Parameter estimation with poisson regression

	Dependent variables								
	N_Vote			N_Thank			N_Gift		
	β	Std	Sig	β	Std	Sig	β	Std	Sig
Constant	4.182	0.009	***	3.312	0.013	***	4.862	0.006	***
s_Doc_Term_Freq	− 0.033	0.004	***	− 0.118	0.005	***	0.006	0.002	*
s_Pat_Term_Freq	0.099	0.003	***	0.134	0.004	***	0.217	0.002	***
Reply_Frequency	0.000	0.000		− 0.004	0.000	***	0.015	0.000	***
Response_Speed	0.019	0.000	***	0.022	0.000	***	0.019	0.000	***
Risk	− 0.053	0.005	***	0.035	0.007	***	− 0.031	0.003	***
Title	0.095	0.002	***	0.135	0.002	***	0.103	0.001	***
s_Doc_Term_Freq× s_Pat_Term_Freq	0.082	0.002	***	0.145	0.003	***	0.084	0.001	***

Significance: *** $p < 0.001$; ** $p < 0.01$; *$p < 0.05$

Thirdly, we checked whether the results were robust to the estimation models. We used OLS and Poisson regression models instead of Negative Binomial regression model. Specifically, we scaled the dependent variables (denoted as s_N_Vote, s_N_Thank and s_N_Gift) in the OLS regression to hold the residual-related assumptions. According to Tables 12 and 13, the results are consistent with those in the main analysis.

Fourthly, we re-conducted the empirical analysis in the case level instead of the doctor level in the main analysis. In the case level, we identified two dependent variables *Symbol_thank* and *Symbol_gift*.[5] *Symbol_thank* = 1 means the patient

---

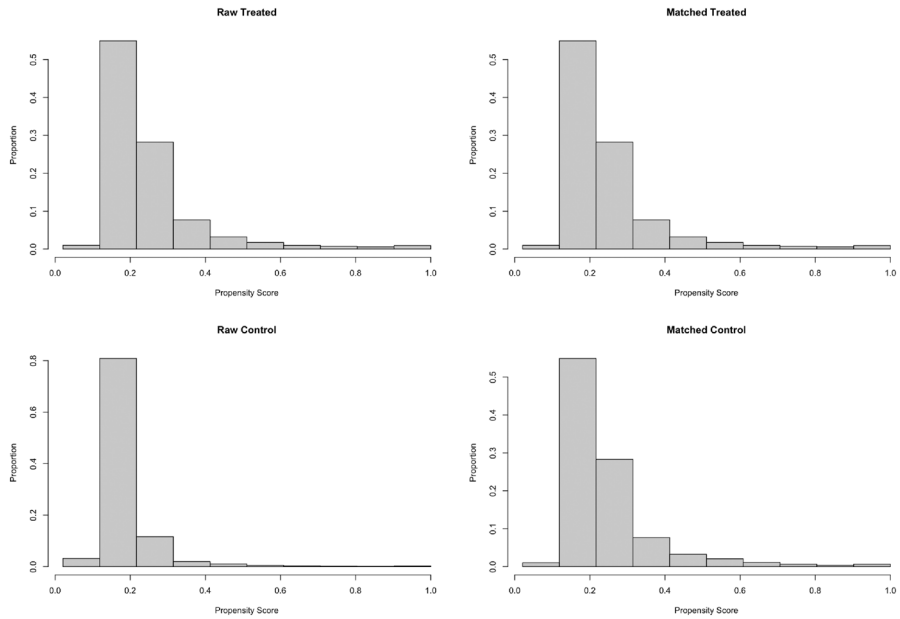[5] Whether a patient voted for a doctor cannot be identified in the case level.

Fig. 6 The distribution of the propensity scores

sent a thank-you letter, and *Symbol_gift* = 1 means the patient sent a digital gift. In addition, we generated an indicator variable *Symbol_satis*, which was calculated as *Max*(*Symbol_thank*, *Symbol_gift*). *Symbol_satis*=1 means the patient provided positive feedback at least in one way. For the independent variables, we identified *Case_Pat_Term_Freq*, *Case_Response_Speed* and *Case_Reply_Frequency*, which were consistent with *Pat_Term_Freq*, *Response_Speed* and *Reply_Frequency* but calculated in the case level.

To control the doctor-related observable and unobservable variables, we added the doctor fixed effect into the model. For patients, since they were highly anonymous for the personal privacy concern, it is impossible to track the patients to get any private information or the other consultation services they received. However, patients may influence the doctors' decision on whether or not to use medical terms. To overcome the selection bias, we used Propensity Score Matching (PSM) to construct a counter-factual framework, which is widely used in economics, statistics, and information systems to address the problem of selection bias [14, 17, 23, 25].

Since this study focuses on the issue whether doctors should use medical terms, we considered the cases where doctors used medical terms as the treatment group, and those where doctors used no medical terms as the control group. We used variables *Case_Pat_Term_Freq*, *Case_Response_Speed*, *Case_Reply_Frequency*, *Response_Speed*, *Case_Reply_Frequency*, *Risk* and *Title* as the matching variables, and then built a Logit model to predict the probability of a doctor used medical terms or not. With the Logit model, propensity scores were calculated for each consultation case, and similar samples in the treatment

Table 14 Parameter estimation with Logistic regression in the case level, with the doctor fixed effect and PSM

	Dependent variables								
	Symbol_thank			Symbol_gift			Symbol_satis		
	β	Std	Sig	β	Std	Sig	β	Std	Sig
Doctor level									
Doctor fixed effect	Yes			Yes			Yes		
Reply_Frequency	− 2.97E−02	1.10E−02	**	− 6.84E+13	2.93E+05	***	− 3.51E−02	9.75E−03	***
Response_Speed	− 8.55E−02	2.41E−02	***	2.18E+11	6.08E+05	***	− 3.75E−02	1.96E−02	.
Risk	1.96E−01	7.35E−02	**	8.10E+13	2.01E+06	***	1.77E−01	6.35E−02	**
Title	3.90E+11	2.93E+11		− 5.39E+14	9.59E+06	***	− 4.19E+10	2.15E+11	
Case level									
Case_Response_Speed	1.19E−04	7.03E−04		6.64E+11	1.76E+04	***	9.96E−04	5.72E−04	.
Case_Reply_Frequency	1.08E−02	6.94E−04	***	1.34E+13	1.67E+04	***	1.54E−02	6.97E−04	***
Case_Doc_Term_Freq	− 3.01E−01	5.90E−02	***	− 5.05E+14	9.91E+05	***	− 3.67E−01	4.59E−02	***
Case_Pat_Term_Freq	− 2.15E−01	5.17E−02	***	− 8.95E+14	1.06E+06	***	− 3.97E−01	4.22E−02	***
Case_Doc_Term_Freq× Case_Pat_Term_Freq	1.64E−01	7.13E−02	*	3.80E+14	1.46E+06	***	2.36E−01	5.36E−02	***

Significance: *** $p < 0.001$; ** $p < 0.01$; *$p < 0.05$; $p < 0.1$

group and the control group could be matched. In this way, the matched samples would have the same potential to enter the treatment/control group. Figure 6 shows the distribution of the propensity scores of the matched samples in the two groups. Finally, with the matched samples, we built Logistic Regression (LR) models to check whether the influence of medical terms used by doctors is robust. Table 14 shows the case-level LR models with the doctor fixed effect and PSM, where the results are consistent with those in the doctor level analysis.;

7 Implications

7.1 Theoretical implications

This paper mainly contributes to three theoretical aspects. Firstly, we empirically and quantitatively explore the influence of using medical terms on service quality with real data, which provides a new semantic perspective for the research of E-health based on text mining. The textual factors in the doctor-patient communication have not been well explored in the literature of E-health. Thus, based on the proposed medical term identification method, some other typical behavior of patients, such as adoption [7, 24, 47, 58], satisfaction and compliments [47, 52], could be re-examined from the semantic perspective of medical terms.

Secondly, the conclusions provide a complement to existing research conducted in the offline situation. Traditional studies unaware of different patients' health literacy hold a negative attitude towards the use of medical terms, which would cause difficulty in understanding and communication for patients [19, 26, 39, 43]. In our study, for patients with limited health literacy, the use of medical terms also has significantly negative influence. However, for patients with high health literacy, the use of medical terms has a significantly positive influence on service quality, which is against the stereotype. The unexpected results derive from the difference between the offline and the online situations could be caused by the special characteristics of E-health in terms of discontinuous online communication, easy access to a large amount of historical communication texts, various forms of online information service tools, etc.

Thirdly, the empirical findings in this study also validate previous theories about E-health. For E-health, the findings validate the positive effect of reply frequency, response speed [57] and individual reputation [31] on service quality. In addition, in the comparison of Tables 8 and 9, the grouped regression results reveal that the doctors' reputation (titles) poses a larger impact on patients with low health literacy than that on patients with high health literacy. Titles are the individual reputation of doctors in the offline situation [31]. According to the signaling theory, the reputation of doctors is a useful cue which could increase the trust of patients [31, 38], especially in E-health where information asymmetry exists. In this way, the difference between patients with high/low health literacy is in line with the theory of novice and expert consumers in marketing. In

product evaluation tasks, novice consumers who are unable to analyze intrinsic cues, tend to use extrinsic cues to make their decisions [5].

7.2 Practical implications

Firstly, the findings of this study provide practical suggestions for both doctors and platforms of E-health. For doctors, using medical terms in an appropriate way could improve their service quality. This suggestion particularly benefits junior doctors without illustrious titles. In addition, when using medical terms, doctors can guide patients to reliable health information websites [35] by taking advantage of time intervals in online communication, to help them understand medical terms. For platforms, the experience shared from other patients and public knowledge articles contributed by doctors should be encouraged and promoted, which may enrich patients' medical knowledge and increase the general health literacy of patients in the E-health community. Besides, platforms could generate health literacy profiles for different patients, and facilitate the judgment of patients' health literacy for doctors. E-health platforms could also develop a recommendation system to match doctors and patients with suitable health literacy. Moreover, E-health platforms are encouraged to develop some online integration tools, such as a specialized search engine for medical terms, to help improve patients' health literacy.

Secondly, we propose a novel deep-learning-based method to identify and extract medical terms from a large scale of doctor-patient communication texts. The proposed method leverages the crowd wisdom of Baidu Wikipedia, which is publicly available and automatically updated. Compared with existing manual-based and dictionary-based methods, the proposed unsupervised method is promising to automatically identify up-to-date medical terms. Such a method provide an effective instrument for E-health platforms. Thus, E-health platforms can apply it to accurately distinguishing doctors who frequently/infrequently use medical terms and patients with high/low health literacy, which is useful for platforms to improve service quality through matching doctors and patients properly.

8 Conclusions

This study explores the influence of medical terms used by doctors on service quality in E-health from the perspectives of doctor-patient communication. To identify and extract medical terms in online doctor-patient communication texts, a novel unsupervised method leveraging the crowd wisdom is proposed with text mining. Based on the identified medical terms, an empirical study is conducted on the data of a Chinese E-health website. The results indicate that for doctors whose patients are of low health literacy, their service quality is negatively affected by the use of medical terms; while for doctors whose patients are of high health literacy, the use of medical terms by doctors significantly improves their service quality in E-health.

There remain some limitations to this study which deserve separated research. Firstly, this study aims to facilitate doctors' decision on whether or not to use

medical terms. Limited to the data availability and privacy concerns, some variables related to the use of medical terms by patients are not traceable, such as income, education and past medical experience. A survey study on the patient health literacy can be conducted in the future work. Secondly, there may exist the self-selection problem on the patient side, which means that unobservable factors of patients would affect their use of medical terms. The problem can be solved by using the patient fixed effect if multiple cases of a patient can be identified, which is unfortunately absent in our dataset. Subject to data privacy, future work can be tried by integrating more data on some other types of platforms, such as online chronic disease communities, where patients can be identified across multiple cases.

Data availability For the data used in the current study, it can be obtained from the corresponding author on reasonable request.

Declarations

Conflict of interest On behalf of all the authors, the corresponding author states that there is no conflict of interest.

References

1. Abedin, B., Milne, D., & Erfani, E. (2020). Attraction, selection, and attrition in online health communities: Initial conversations and their association with subsequent activity levels. *International Journal of Medical Informatics, 141*, 104216. https://doi.org/10.1016/j.ijmedinf.2020.104216
2. Atanasova, S., Kamin, T., & Petrič, G. (2018). The benefits and challenges of online professional-patient interaction: Comparing views between users and health professional moderators in an online health community. *Computers in Human Behavior, 83*, 106–118. https://doi.org/10.1016/j.chb.2018.01.031
3. Bao, S., Li, R., Yu, Y., & Cao, Y. (2008). Competitor mining with the web. *IEEE Transactions on Knowledge and Data Engineering, 20*(10), 1297–1310. https://doi.org/10.1109/TKDE.2008.98
4. Berkman, N. D., Davis, T. C., & McCormack, L. (2010). Health literacy: What is it? *Journal of Health Communication, 15*(sup2), 9–19. https://doi.org/10.1080/10810730.2010.499985
5. Biswas, A., & Sherrell, D. L. (1993). The influence of product knowledge and brand name on internal price standards and confidence. *Psychology & Marketing, 10*(1), 31–46. https://doi.org/10.1002/mar.4220100104
6. Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data*. Cambridge University Press.
7. Cao, X., Liu, Y., Zhu, Z., Hu, J., & Chen, X. (2017). Online selection of a physician by patients: Empirical study from elaboration likelihood perspective. *Computers in Human Behavior, 73*, 403–412. https://doi.org/10.1016/j.chb.2017.03.060
8. Cerqueus, T., Bonnaud, J., Dashkov, O., & Morin, E. (2020). Thesaurus matching in electronic commerce. *Electronic Commerce Research*. https://doi.org/10.1007/s10660-020-09438-9
9. Chau, M., Li, T. M., Wong, P. W., Xu, J. J., Yip, P. S., & Chen, H. (2020). Finding people with emotional distress in online social media: A design combining machine learning and rule-based classification. *MIS Quarterly, 44*(2), 933–955. https://doi.org/10.25300/MISQ/2020/14110

10. Chen, L., Baird, A., & Straub, D. (2019). Fostering participant health knowledge and attitudes: An econometric study of a chronic disease-focused online health community. *Journal of Management Information Systems, 36*(1), 194–229. https://doi.org/10.1080/07421222.2018.1550547

11. Chen, L., Baird, A., & Straub, D. (2020). A linguistic signaling model of social support exchange in online health communities. *Decision Support Systems, 130*, 113233. https://doi.org/10.1016/j.dss.2019.113233

12. Chen, S., Guo, X., Wu, T., & Ju, X. (2020). Exploring the online doctor-patient interaction on patient satisfaction based on text mining and empirical analysis. *Information Processing & Management, 57*(5), 102253. https://doi.org/10.1016/j.ipm.2020.102253

13. Chen, W., Chen, J., & Yin, G. (2020). Exploring side effects of ridesharing services in urban china: role of pollution-averting behavior. *Electronic Commerce Research*. https://doi.org/10.1007/s10660-020-09443-y

14. Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics, 84*(1), 151–161.

15. Deng, S., Huang, Z. J., Sinha, A. P., & Zhao, H. (2018). The interaction between microblog sentiment and stock returns: An empirical examination. *MIS Quarterly, 42*(3), 895–918. https://doi.org/10.25300/MISQ/2018/14268

16. Huang, K., Chengalur-Smith, I. (2014). A social capital perspective to understand individual contribution of social support in healthcare virtual support communities. In *2014 47th Hawaii international conference on system sciences*, pp. 3489–3498 https://doi.org/10.1109/HICSS.2014.435.

17. Jabr, W., Mookerjee, R., Tan, Y., & Mookerjee, V. S. (2014). Leveraging philanthropic behavior for customer support: The case of user support forums. *MIS Quarterly, 38*(1), 187–208.

18. Kleiber, C., & Zeileis, A. (2008). *Applied econometrics with R*. Springer.

19. Korsch, B. M., Gozzi, E. K., & Francis, V. (1968). Gaps in doctor-patient communication: I–Doctor-patient interaction and patient satisfaction. *Pediatrics, 42*(5), 855–871.

20. Le, Q., Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*. pp. 1188–1196.

21. Lehtinen, U., & Lehtinen, J. R. (1991). Two approaches to service quality dimensions. *The Service Industries Journal, 11*(3), 287–303. https://doi.org/10.1080/02642069100000047

22. Lewis, R. C., & Booms, B. H. (1983). The marketing aspects of service quality. *Emerging Perspectives Services Marketing, 65*(4), 99–107.

23. Li, X.: The impact of place-of-origin on price premium for agricultural products: Empirical evidence from taobao. com. Electronic Commerce Research pp. 1–24 (2020)

24. Li, Y., Song, Y., Zhao, W., Guo, X., Ju, X., & Vogel, D. (2019). Exploring the role of online health community information in patients decisions to switch from online to offline medical services. *International Journal of Medical Informatics, 130*, 103951. https://doi.org/10.1016/j.ijmedinf.2019.08.011

25. Li, Z., Song, Y., & Xu, X. (2019). Incorporating facial attractiveness in photos for online dating recommendation. *Electronic Commerce Research, 19*(2), 285–310.

26. Links, A., Callon, W., Wasserman, C., Walsh, J., Beach, M., & Boss, E. (2019). Surgeon use of medical jargon with parents in the outpatient setting. *Patient Education and Counseling, 102*(6), 1111–1118. https://doi.org/10.1016/j.pec.2019.02.002

27. Liu, F.T., Ting, K.M., Zhou, Z.H. (2008). Isolation forest. In *2008 Eighth IEEE international conference on data mining*. pp. 413–422. IEEE.

28. Liu, N., Tong, Y., & Chan, H. C. (2020). Dual effects of social support seeking in patient-centric online healthcare communities: A longitudinal study. *Information & Management, 57*(8), 103270. https://doi.org/10.1016/j.im.2020.103270

29. Liu, S., Xiao, W., Fang, C., Zhang, X., & Lin, J. (2020). Social support, belongingness, and value co-creation behaviors in online health communities. *Telematics and Informatics, 50*, 101398. https://doi.org/10.1016/j.tele.2020.101398

30. Liu, S., Zhang, M., Gao, B., & Jiang, G. (2020). Physician voice characteristics and patient satisfaction in online health consultation. *Information & Management, 57*(5), 103233. https://doi.org/10.1016/j.im.2019.103233

31. Liu, X., Guo, X., Wu, H., & Wu, T. (2016). The impact of individual and organizational reputation on physicians appointments online. *International Journal of Electronic Commerce, 20*(4), 551–577. https://doi.org/10.1080/10864415.2016.1171977

32. Liu, Y., Ren, C., Shi, D., Li, K., & Zhang, X. (2020). Evaluating the social value of online health information for third-party patients: Is uncertainty always bad? *Information Processing & Management, 57*(5), 102259. https://doi.org/10.1016/j.ipm.2020.102259

33. Locke, M. R., Shiyanbola, O. O., & Gripentrog, E. (2014). Improving prescription auxiliary labels to increase patient understanding. *Journal of the American Pharmacists Association, 54*(3), 267–274. https://doi.org/10.1331/JAPhA.2014.13163

34. Masland, M. C., Kang, S. H., & Ma, Y. (2011). Association between limited english proficiency and understanding prescription labels among five ethnic groups in california. *Ethnicity & Health, 16*(2), 125–144. https://doi.org/10.1080/13557858.2010.543950

35. McMullan, M. (2006). Patients using the internet to obtain health information: How this affects the patient-health professional relationship. *Patient Education and Counseling, 63*(1), 24–28. https://doi.org/10.1016/j.pec.2005.10.006

36. Mirzaei, T., & Esmaeilzadeh, P. (2021). Engagement in online health communities: Channel expansion and social exchanges. *Information & Management, 58*(1), 103404. https://doi.org/10.1016/j.im.2020.103404

37. Mirzaei, T., & Kashian, N. (2020). Revisiting effective communication between patients and physicians: Cross-sectional questionnaire study comparing text-based electronic versus face-to-face communication. *Journal of Medicle Internet Research, 22*(5), e16965. https://doi.org/10.2196/16965

38. Mou, J., Shin, D. H., & Cohen, J. F. (2017). Trust and risk in consumer acceptance of e-services. *Electronic Commerce Research, 17*(2), 255–288.

39. O'Connell, R., Hartridge-Lambert, S., Din, N., St John, E., Hitchins, C., & Johnson, T. (2013). Patients understanding of medical terminology used in the breast clinic. *The Breast, 22*(5), 836–838. https://doi.org/10.1016/j.breast.2013.02.019

40. Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing, 49*(4), 41–50.

41. Pee, L. G. (2016). Customer co-creation in b2c e-commerce: Does it lead to better new products? *Electronic Commerce Research, 16*(2), 217–243.

42. Pieterse, A. H., Jager, N. A., Smets, E. M., & Henselmans, I. (2013). Lay understanding of common medical terminology in oncology. *Psycho-Oncology, 22*(5), 1186–1191. https://doi.org/10.1002/pon.3096

43. Schnitzler, L., Smith, S. K., Shepherd, H. L., Shaw, J., Dong, S., Carpenter, D. M., Nguyen, F., & Dhillon, H. M. (2017). Communication during radiation therapy education sessions: The role of medical jargon and emotional support in clarifying patient confusion. *Patient Education and Counseling, 100*(1), 112–120. https://doi.org/10.1016/j.pec.2016.08.006

44. Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., & Platt, J. C. (2000). Support vector method for novelty detection. *Advances in Neural Information Processing Systems, 12*, 582–588.

45. Shang, Y., Liu, J. (2016). Health literacy: Exploring health knowledge transfer in online healthcare communities. In *2016 49th Hawaii International conference on system sciences (HICSS)*, pp. 3143–3151. https://doi.org/10.1109/HICSS.2016.395.

46. Sharma, S. (1996). *Applied multivariate techniques*. Wiley.

47. Shim, M., & Jo, H. S. (2020). What quality factors matter in enhancing the perceived benefits of online health information sites? application of the updated delone and mclean information systems success model. *International Journal of Medical Informatics, 137*, 104093. https://doi.org/10.1016/j.ijmedinf.2020.104093

48. Smith, S. K., Dixon, A., Trevena, L., Nutbeam, D., & McCaffery, K. J. (2009). Exploring patient involvement in healthcare decision making across different education and functional health literacy groups. *Social Science & Medicine, 69*(12), 1805–1812. https://doi.org/10.1016/j.socscimed.2009.09.056

49. Swar, B., Hameed, T., & Reychav, I. (2017). Information overload, psychological ill-being, and behavioral intention to continue online healthcare information search. *Computers in Human Behavior, 70*, 416–425. https://doi.org/10.1016/j.chb.2016.12.068

50. Voorhees, E. M., & Harman, D. K. (2005). *TREC experiment and evaluation in information retrieval*. MIT press.

51. Wu, H., & Deng, Z. (2019). Knowledge collaboration among physicians in online health communities: A transactive memory perspective. *International Journal of Information Management, 49*, 13–33. https://doi.org/10.1016/j.ijinfomgt.2019.01.003

52. Wu, H., Deng, Z., Wang, B., & Wu, T. (2020). Online service qualities in the multistage process and patients compliments: A transaction cycle perspective. *Information & Management, 57*(5), 103230. https://doi.org/10.1016/j.im.2019.103230

53. Wu, H., & Lu, N. (2018). Service provision, pricing, and patient satisfaction in online health communities. *International Journal of Medical Informatics, 110*, 77–89. https://doi.org/10.1016/j.ijmedinf.2017.11.009

54. Xia, H., Yang, Y., Pan, X., Zhang, Z., & An, W. (2020). Sentiment analysis for online reviews using conditional random fields and support vector machines. *Electronic Commerce Research, 20*(2), 343–360.

55. Xiao, N., Sharman, R., Rao, H., & Upadhyaya, S. (2014). Factors influencing online health information search: An empirical analysis of a national cancer-related survey. *Decision Support Systems, 57*, 417–427.

56. Yan, L., & Tan, Y. (2014). Feeling blue? go online: An empirical study of social support among patients. *Information Systems Research, 25*(4), 690–709. https://doi.org/10.1287/isre.2014.0538

57. Yang, H., Guo, X., & Wu, T. (2015). Exploring the influence of the online physician service delivery process on patient satisfaction. *Decision Support Systems, 78*, 113–121. https://doi.org/10.1016/j.dss.2015.05.006

58. Yang, H., Guo, X., Wu, T., & Ju, X. (2015). Exploring the effects of patient-generated and system-generated information on patients online search, evaluation and decision. *Electronic Commerce Research and Applications, 14*(3), 192–203. https://doi.org/10.1016/j.elerap.2015.04.001

59. Zhang, J., Zhang, J., & Zhang, M. (2019). From free to paid: Customer expertise and customer satisfaction on knowledge payment platforms. *Decision Support Systems, 127*, 113140. https://doi.org/10.1016/j.dss.2019.113140

60. Zhang, X., Guo, F., Xu, T., & Li, Y. (2020). What motivates physicians to share free health information on online health platforms? *Information Processing & Management, 57*(2), 102166. https://doi.org/10.1016/j.ipm.2019.102166

61. Zhang, X., Guo, X., Hung Lai, K., & Yi, W. (2019). How does online interactional unfairness matter for patient-doctor relationship quality in online health consultation? the contingencies of professional seniority and disease severity. *European Journal of Information Systems, 28*(3), 336–354.

62. Zhao, H., Fu, S., & Chen, X. (2020). Promoting users intention to share online health articles on social media: The role of confirmation bias. *Information Processing & Management, 57*(6), 102354. https://doi.org/10.1016/j.ipm.2020.102354

Authors and Affiliations

Jilong Zhang[1] · Jin Zhang[1] · Kanliang Wang[1] · Wei Yan[2,3]

Jilong Zhang
zhangjilong@ruc.edu.cn

Kanliang Wang
klwang@ruc.edu.cn

Wei Yan
yanwei@cuc.edu.cn

[1]   Department of Management Science and Engineering, School of Business, Renmin University of China, Beijing 100872, China

[2]   State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China

[3]   School of Economics and Management, Communication University of China, Beijing 100024, China