MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-810

Consumer Health Information Needs and Question Classification: Analysis of Hypertension Related Questions Asked by Consumers on a Chinese Health Website

Haihong Guo[a], Jiao Li[a], Tao Dai[a]

[a] *Institute of Medical Information & Library, Chinese Academy of Medical Sciences, Beijing, China*

Abstract

This study built up a classification schema of consumer health questions which consisted of 48 quaternary categories and 35 annotation rules. Using such a schema, we manually classified 2,000 questions randomly selected from nearly 100 thousand hypertension-related messages posted by consumers on a Chinese health website to analyze the information needs of health consumers. The results showed questions in the categories of treatment, diagnosis, healthy lifestyle, management, epidemiology, and health provider choosing were 48.1%, 23.8%, 11.9%, 5.2%, 9.0%, and 1.9% respectively. The comparison of the questions asked by consumers and physicians showed that their health information needs were significantly different (P<0.0001).

Keywords:
Health Consumers; Information Needs; Hypertension; Question Classification; Consumer Health Informatics.

Introduction

The Internet is increasingly becoming a main resource for consumers to acquire health information. 80% of internet users (i.e. 59% of all adults) in the U.S. have looked online for health information [1]. In China, health channels of main portals and professional health websites have become one of the main resources for health consumers [2], and the number of internet users has increased to about 632 million in 2014 [3]. Many researches have proven that health related information online could impact consumers' health decisions and behaviors [1,4]. Health websites that target specific information needs are burgeoning in response to the high demand and impact [5]. However, making health information available does not necessarily guarantee its accessibility and usability [6]. Consumers have difficulty in expressing their information needs using accurate medical query terms, and further failed to retrieve relevant health information [78]. Thus, it is crucial to develop a method to identify health information needs from questions asked by consumers.

In previous studies, a series of templates were developed to guide question composition, and identify meaningful concepts and their relationships, which were further used to construct query strategy [9]. This method requires consumers to follow the scheme when asking questions, which would affect its usability. Some studies used metathesaurus (such as UMLS and the on-going Consumer Health Vocabulary) [10] to assign multi topics of questions so as to construct navigational exploration interface or generate semantic query strategy [11]. It is hard to specify the information needs (concerned aspect), although the topic has been identified. For example, though it could identify that a question was about hypertension, it was difficult to distinguish whether the user wanted to know its diagnosis, treatment, prognosis, or diet, for eaxmple. A promising way to overcome the shortage is to build up a question classification schema to define question topic and its concerned aspects [12].

For professional health-related questions, several classification schemas have been developed, such as the International Classification of Primary Care [13] and the Taxonomies of Generic Clinical Questions (TGCQ) [14,15]. They have proven to be useful when analyzing physicians' and case managers' information needs [15,16] but not suitable for consumers' questions due to the diffference of information needs between physicians and consumers [17,18]. A Layered Model of Context for Consumer Health Information Searching (LMCC) intended to describe consumers' interest topics on cognitive layers [19] but not define the classficassion schema in a systematic manner. In this study, we aim to build up a consumer health question classification schema to understand and specify users' informational needs.

To test the usefulness of classification schema, we used hypertension related questions asked by Chinese consumers as a test dataset. Hypertension has become the main risk factor of over half of cardiovascular diseases, such as stroke and coronary heart disease. There were 270 million patients (that is, at least 2 patients out of 10 alduts) with hypertension in China in 2012, and the number has continued to increase at a rate of 3.1% per year [20]. Thus, hypertension-related questions have become more frequently asked with large variability on the Internet.

Materials and Methods

Data Collection

Messages posted by health consumers from 01 January to 10 August 2014 with a tag of "hypertension (高血压)" or "blood pressure (血压)" under the Q&A (有问必答) section on xywy.com (http://www. xywy.com/, a Chinese health website with over 35 million users) were collected and imported into a MySQL database. The resulting database included 98,032 messages. To conduct an in-depth analysis of the questions contained in the messages, we randomly selected 2,000 messages.

"Question" is defined as a request that a consumer posted to the website on a certain subject to seek answers from professionals, which was identified based on meaning, not form. This study was focused on questions related to

hypertension (高血压), which was sometimes expressed as "high blood pressure (高血压)," or simply as "high pressure (高压)." So, we manually discarded messages that did not match the definition and that were irrelevant to hypertension, but with the similar words such as "high pressure oxygen (高压氧)," "hyperbaric cabin (高压舱)," "high voltage(高压电)," "pressure cooker (高压锅)," etc.. Another new message was randomly selected from the database when an irrelevant message was discarded from the sample, so as to keep the sample size at 2,000.

Classification Schema Development

A topic-based classification schema was developed based on TGCQ [14,15] and LMCC [19]. The categories of clinical related questions (including diagnosis, treatment, management, epidemiology, and their narrower terms) were mainly selected from TGCQ, while the categories of non-clinical questions (such as healthy lifestyle and health provider choosing, and their narrower terms) were mainly selected and expanded from LMCC. We divided some categories into more specific sub-categories so as to code the specific information needs. For instance, the diet category, under healthy lifestyle, was further divided into five tertiary categories, including how to eat, food choosing, interactions, action mechanism, and general.

One of the authors (specialized in medical informatics), classified all the 2,000 sample questions following the classification schema. During the manual classification progress, some categories were added to accommodate questions that did not fall into any existing medical or non-medical specialty. We developed a list of annotation rules and enumerated some general question types for each of the smallest category, so as to improve consistency among coders and the usability of the classification. For example, questions as the following types were coded as 1.1.1.1 (diagnosis→interpretation of clinical finding→symptom):

- I have (or somebody else has) symptom x, what's the condition / matter? (我有（某人有）症状X，是什么情况？ / 是怎么回事？)
- What cased symptom x? (是什么引起症状X？)
- What's the cause of symptom x? (症状X的原因是什么？)

The website (xywy.com) provides a template for users to generate question including three parts: (1) *describe your health status (病情描述)*, (2) *treatments or tests in the past (曾经的治疗或检查情况)*, and (3) *what kinds of help do you want (想得到怎样的帮助).* This template might lead to consumers' confusion on question fill-up. To deal with this case, we developed a rule: if there is the phrase "what kinds of help do you want (想得到怎样的帮助)" in the message, then code the first question after the phrase as the main topic, and successively code the questions followed by as minor topics. Otherwise, we coded the first question in the message as the main topic.

In this way, we developed the preliminary classification schema of consumer health questions, which had 101 topic categories and 32 annotation rules. The classification schema and questions coding was then modified by the following steps:

Firstly, four volunteers (two with medical education backgrouds, the others with informatics backgroud) used the classification to independently code 200 questions randomly selected from the sample, and each volunter made suggestions to specify the rules and increase some categories to accommodate the questions. The author compared the consistency of the five coding results (including the result of herself), and categorised the 200 questions into three groups: all annotators agreed on the topic (n=73), only one annotator disagreed (n=63), and more than one annotator disagreed (n=64). Then we focused on the last group, looking for problematic and ambiguous questions. Analysis of these inconsistencies allowed us to address ambiguous elements in the classification via specifying annotation rules and changing the description of the example general question types.

Secondly, the revised classification was distributed to the five annotators who independently annotated another 300 questions randomly selected from the remaining 1,800 samples. This step was done to measure the interrater reliability of the revised classification schema, as well as to modify it further.

Lastly, the three volunteers annotated the remaining 1,500 messages. Each of them annotaed 500 independently to ensure all of the 2,000 sample messages were annotated by at least two annotators. The codes agreed upon by this step were regarded as the final ones. Then we calculated the number of questions in each topic category, and deleted categories that did not have any questions filled in (e.g. physical characteristics of drugs, pharmacodynamics, mechanism of drug action).

Statistical Analysis

Descriptive analysis was used to calculate the frequency of question topics (main topic only, and all topics respectively). The kappa=$(P_o-P_e)/(1-P_e)$ statistic, which could correct agreement that occured by chance, was used to determine the interrater reliability of the question classification, where P_o was the observed agreement and P_e was the agreement expected by chance. When the number of categories was large, as in this study, P_e would be close to zero, and the kappa value would be close to P_o. Thus we directly used P_o as kappa value. The bigger the kappa value, the better the agreement. We supposed that when the user asked more than one question, it was acceptable to answer any one of them. Therefore, a liberal reliability criteria was used; a match was recorded if either the main or minor topics assigned by one annotator matched the other's assignment.

We merged the topic classification of consumer health questions developed in this study and the Taxonomy of Generic Clinical Questions (TGCQ) [15] into one classification table, and then used a chi-square test to compare the frequency distributions of topics asked by consumers and professionals.

Results

Classification of Consumer Health Questions

The final classification schema was a four hierarchical levels of specificity, consisted of 48 quaternary categories (Table 1), and included 35 annotation rules, down from 101 categories in the preliminary version. The first level included seven broad areas: diagnosis, treatment, management, epidemiology, healthy lifestyle, health provider choosing, and other. Condition/finding management questions asked what steps to take without distinguishing between diagnostic steps and

therapeutic steps [15]. To answer them, one should first give a diagnosis, and then the suggestion of treatment. A branching structure of secondary, tertiary, and quaternary levels describes more and more specific topics of the questions. One or more closely related generic questions were listed for each quaternary category. For instance, the question "A 65-year-old man with unsteady high blood pressure, what's the best blood pressure drug to eat? (65岁老人血压高经常不稳定，吃哪种降压药最好?)" would be coded as 2.1.2.1 (treatment→drug therapy→efficacy/ indications→treatment), and the generic question type could be "Condition y, what's the best drug (to eat / take / use)?"

Table 1 – Classification of consumer health questions.

Code	Primary	Secondary	Tertiary	Quaternary	Frequency(%) (main code)	Frequency(%) (all codes)
1.1.1.1	diagnosis	interpretation of symptom			39(2.0)	44(1.7)
1.1.2.1		clinical finding	sign		146(7.3)	152(5.8)
1.1.3.1			test finding		7(0.4)	8(0.3)
1.1.4.1			multiple findings		286(14.3)	302(11.6)
1.2.1.1		criteria			35(1.8)	37(1.4)
1.3.1.1		test	indications/ efficacy		36(1.8)	55(2.1)
1.3.2.1			accuracy		4(0.2)	6(0.2)
1.3.3.1			timing		6(0.3)	6(0.2)
1.3.4.1			method		4(0.2)	5(0.2)
1.4.1.1		orientation	condition		4(0.2)	5(0.2)
1.5.1.1		cost			0(0.0)	2(0.1)
2.1.1.1	treatment	drug therapy	how to use	general	4(0.2)	7(0.3)
2.1.1.2				dosage	8(0.4)	11(0.4)
2.1.1.3				timing	38(1.9)	52(2.0)
2.1.2.1			efficacy/ indications	treatment	324(16.2)	389(14.9)
2.1.2.2				prevention	3(0.2)	7(0.3)
2.1.3.1			adverse effects	caused by drug	29(1.5)	46(1.8)
2.1.3.2				control	2(0.1)	5(0.2)
2.1.3.3				safety/contraindications	23(1.2)	27(1.0)
2.1.4.1			interactions		20(1.0)	25(1.0)
2.1.5.1			name		1(0.1)	2(0.1)
2.1.6.1			cost		1(0.1)	1(0.0)
2.1.7.1			availability		1(0.1)	1(0.0)
2.1.8.1			brand		2(0.1)	3(0.1)
2.2.1.1		not limited to but	efficacy/ indications	treatment	473(23.7)	621(23.8)
2.2.1.2		may include drug		prevention	7(0.4)	16(0.6)
2.2.2.1		therapy	timing		4(0.2)	8(0.3)
2.2.3.1			how to do it		1(0.1)	1(0.0)
2.2.4.1			safety/ contra/ sequelae		12(0.6)	26(1.0)
2.2.5.1			cost		2(0.1)	8(0.3)
3.1.1.1	management	condition/ finding			114(5.7)	136(5.2)
4.1.1.1	epidemiology	prevalence			0(0.0)	1(0.0)
4.2.1.1		etiology	causation/ association	risk factors	111(5.6)	149(5.7)
4.2.1.2				genetics	3(0.2)	4(0.2)
4.3.1.1		prognosis			51(2.6)	82(3.1)
5.1.1.1	healthy	diet	how to eat		4(0.2)	4(0.2)
5.1.2.1	lifestyle		food choosing	efficacy	69(3.5)	97(3.7)
5.1.2.2				contraindications	29(1.5)	40(1.5)
5.1.3.1			interactions		2(0.1)	4(0.2)
5.1.4.1			general		19(1.0)	32(1.2)
5.2.1.1		exercise			7(0.4)	15(0.6)
5.3.1.1		weight-losing			3(0.2)	3(0.1)
5.4.1.1		mood control			2(0.1)	3(0.1)
5.5.1.1		general			35(1.8)	107(4.1)
6.1.1.1	health provider	hospital			10(0.5)	18(0.7)
6.2.1.1	choosing	department			11(0.6)	25(1.0)
6.3.1.1		doctor			3(0.2)	6(0.2)
7.1.1.1	other				5(0.3)	6(0.2)
Total					2000(100)	2610(100)

General Topics of Questions Asked by Health Consumers

The 2,000 sample messages were coded with 2,000 main codes and 610 minor codes, the frequency of each topic category was shown in Table 1. 48% of the questions were asked about treatment, which indicated that nearly half of the health consumers posting questions on the website have noticed that they or somebody they care about has had some health problem and needed to be treated. Almost half (45.9%) of the treatment questions were referred in particular to drug

therapy, including how to use drugs (5.6%), how to choose drugs for a particular condition (31.5%) and adverse effects of drugs (6.2%). 23.8% of the questions were asked about diagnosis, and the majority (19.4%) were seeking interpretation of consumers' specific clinical findings in reality as each pertained to the symptom they felt (1.7%), the sign they knew from physical examination (5.8%), or multiple kinds of findings they got (11.6%). 5.2% of the questions were coded as 3.1.1.1 (management of condition or findings) because they were not specified in diagnosis or treatment, and more than half of them (54.4%) were just enumerated as a series of clinical findings without any interrogative sentence or term.

11.9% of questions were asking what to do or what not to do in everyday life in order to keep healthy or get well from certain illnesses. More than half (58.4%) of them were concerned with diet or nourishment. Among the 9.0% of epidemiological questions, 5.7% were about risk factors, both risk factors of the diseases they had and if their condition would be harmful to some particular conditions, e.g., pregnancy, parturition and sexual life. 3.1% of questions were about prognosis and we thought many of them were mainly expressing anxiety as the asker wanted to get an affirmative reply to allay their worry [19]. Among the 1.9% provider choosing questions, half were about medical department choosing for specific conditions or clinical findings, which indicates that medical guide service would be a promising area for health websites.

Interrater Reliability

The kappa statistic for the five coders was 0.63 in the quaternary level of the classification, indicating "substantial" reliability. When just the the primary and secondary levels were considered, the kappa value increased to 0.75. When only the seven broad areas in the primary level were considered, agreement was almost perfect (kappa=0.82).

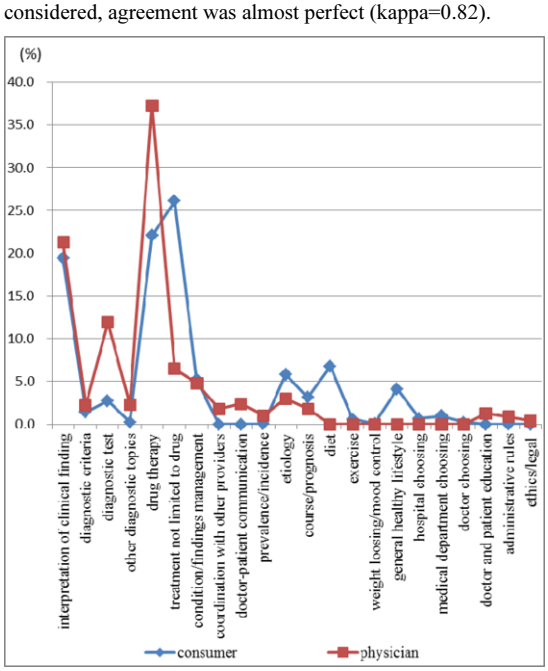

Figure 1 – The frequency distribution of the secondary topics of the questions asked by health consumers and physicians.

Difference of Health Information Needs Between Consumers and Physicians

The chi-square test of the topic distributions of the questions asked by the two groups respectively showed that health information needs of consumers were significantly different from clinical information needs of physicians. The pearson chi-square value on the quaternary category was 1477.89 ($P<0.0001$), and was 854.38 ($P<0.0001$) on the secondary category. Figure 1 shows the frequency difference along the secondary topics of the questions between health consumers and physicians. For exmaple, phyicians are more interested in the diagnosis test than consumers (11.9% vs. 2.8%).

Discussion

Health consumers and physicians both asked questions about diagnosis, treatment, management of conditions and findings, and epidemiology. Besides these topics, health consumers also asked how to keep healthy or help recovery in daily life, because many of them recognized that lifestyle, such as diet, exercise, weight loosing, and mood control, would impact their health status as well [21]. While physicians seldom asked these questions during a patient encounter, it might be because they mainly focused on medical service rather than lifestyle advice [18]. Similarly, health consumers never asked questions about coordination with other providers, doctor-patient communication, doctor and patient education, administrative rules, ethics, and legal issues, since these tasks were usually regarded as health providers' responsibility.

Both of the two groups sought answers for the interpretation of clinical findings, while the questions posted by health consumers were much more vague, the frequency of questions with multiple findings were two times more than that of physician-based inquiries. It might be because they could not distinguish which findings were most important, so they tended to put all the findings they knew. Physicians were more concerned ahout what test to choose for a particular situation and when or how to do it (the question frequency was three times more than that of health consumers), because they wanted to know how to diagnose [14,15], while health consumers wanted to know the diagnosis. Though the frequency of treatment questions was almost equal in the two groups (48.2% vs. 43.7%), physicians' questions were more specified to drug therapy (37.2% vs. 22.1%), and they sometimes asked those questions on very specialized sides, such as composition, pharmacodynamics, action mechanism, and serum levels of drugs, which were rarely asked by health consumers.

Health consumers were mainly concerned about what was wrong with their health (or the health of someone they care about), what went wrong, how to treat it (including choosing which provider to treat), possible adverse drug effects, cross interactions or dangers with other conditions (e.g. pregnancy, breast feeding, etc.), duration of recovery from the illness, and health maintenance in everyday life (e.g. dietary suggestions). Thus, they seldom asked questions that were commonly regarded as the physicians' tasks or too medical specialized.

Conclusion

This study built a classification schema of consumer health questions which consisted of 48 quaternary categories and 35 annotation rules. Five annotators followed this schema and

classified 2,000 questions randomly selected from nearly 100,000 hypertension-related messages posted by consumers on a Chinese health website. The potential uses of this study were identified as follows. First, the classification could be used to organise large collections of consumer health questions, so as to improve retrieval efficiency. Second, the distribution of question topics could be used to guide the building of a knowledge base for health websites, such as setting priorities to building a knowledge base for those frequently asked questions. Third, the coded questions could be used as a corpus for studies, such as training machines to automatically classify the topics of questions posted by health consumers, and further used for the monitoring of hot health topics and automatically generating answers. Last, but not least, the perceived information needs of health consumers could be used to help set the priorities of medical research and patient education.

This study also had some limitations, the sample questions were collected from only one website and defined to be hypertension or blood pressure related; thus, the applicability of the classification on other settings has yet to be studied further. The information needs analyzed in this study were "user based," that is, the topics were assigned without considering the best way to answer it. Although we achieved substantial interrater reliability, surpassed several similar research, such as assigning topics to generic clinical questions (kappa=0.53) [15], and assigning medical subject headings and subheadings (MeSH terms) to journal articles (consistency percentages was 0.43%) [22], the classification and annotation rules have yet to be modified and tested with larger and more diverse sample questions.

Acknowledgments

This research is supported by the National Key Technology Research and Development Program of China (Grant No. 2013BAI06B01). The authors would like to thank Dr. Yueping Sun, Hongyan Liu, Xiaolin He, Ze Zhang, Hui Wang at IMICAMS for their helpful suggestions on data processing.

References

[1] Fox S. The Social Life of Health Information. Washington, DC: Pew Internet & American Life Project. 2011.

[2] Links-group, Health Sohu. 2009 China Health Communication and Popularization Survey. 2009.

[3] China Internet Network Information Center. China Internet Network Development State Statistic Report. 2014.7.

[4] Ying Z. The Causal Relationship Between Health Related Internet Use and Health Behaviors. Wuhan: Huazhong University of Science and Technology. 2011.

[5] Dai T. Enriching online access to health information for the public: efforts from a medical library in China. Chinese Journal of Library and Information Science 2013: 6(2): 1-13.

[6] Zhang Y. A Review of Search Interfaces in Consumer Health Websites. 2011. [Online: https://www.ischool.utexas.edu/~yanz/HCIR2011_Zhang.pdf; Accessed Nov 2014].

[7] Gunther E, and Christian K. How Do Consumers Search for and Appraise Health Information on the World Wide Web? Qualitative Study Using Focus Groups, Usability Tests, and in-Depth Interviews. British Medical Journal 2002: 324(7337): 573-577.

[8] Zeng QT, Crowell J, Plovnick RM, et al. Assisting Consumer Health Information Retrieval with Query Recommendations. Journal of the American Medical Informatics Association 2006: 13(1): 80–90.

[9] Hallett C, Power R, and Scott D. Composing Questions through Conceptual Authoring. Computational Linguistics 2007: 33(1): 105-133.

[10] Qing TZ, and Tony T. Open Source and Collaborative Development of Consumer Health Vocabulary. [online: http://consumerhealthvocab.org/; accessed Nov 2014].

[11] Licong C, Rong X, Zhihui L, et al. Multi-topic Assignment for Exploratory Navigation of Consumer Health Information in NetWellness using formal concept analysis. BMC Medical Informatics and Decision Making 2014: 14: 63-75.

[12] Cao YG, Cimino JJ, Ely J, et al. Automatically Extracting Information Needs from Complex Clinical Questions. Journal of Biomedical Informatics 2010: 43(6): 962-971.

[13] Gebel RS, and Okkes IM. International Classification of Primary Care (ICPC-2-NL). Utrecht: Nederlands Huisartsen Genootschap, 2000.

[14] Ely JW, Osheroff JA, Ebell MH, et al. Analysis of Questions Asked by Family Doctors Regarding Patient Care. BMJ 1999: 319(7206): 358-61.

[15] Ely JW, Osheroff JA, Gorman PN, et al. A Taxonomy of Generic Clinical Questions: Classification Study. BMJ 2000: 321(7258): 429-32.

[16] Schnall R, Cimino JJ, Currie LM, and Bakken S. Information needs of case managers caring for persons living with HIV. J Am Med Inform Assoc 2011: 18(3): 305-8.

[17] Cécile RLB, and Frans JM. Classifying Health Questions Asked by the Public Using the ICPC-2 Classification and a Taxonomy of Generic Clinical Questions: An Empirical Exploration of the Feasibility. Health Communication 2010: 25(2): 175-181.

[18] Reeder B, Le T, Thompson HJ, et al. Comparing Information Needs of Health Care Providers and Older Adults: Findings from a Wellness Study. MedInfo2013: Proceedings of the 14th World Congress on Medical and Health Informatics, 2013, 192: 18-22.

[19] Yan Z. Toward a Layered Model of Context for Health Information Searching: An Analysis of Consumer-Generated Questions. Journal of the American Society for Information Science and Technology 2013: 64(6): 1158-1172.

[20] National Center for Cardiovascular Diseases, China. Report on Cardiovascular Diseases in China (2013). Beijing: Encyclopedia of China Publishing House, 2014.4.

[21] Publicity Department of National Health and Family Planning Commission of the People's Republic of China, Chinese Health Education Center. 2012 Chinese Residents Health Literacy Monitoring Report. 2013:6-12.

[22] Funk ME, and Reid CA. Indexing consistency in MEDLINE. Bull Med Libr Assoc 1983: 71: 176-83.

Address for correspondence

Tao Dai, Institute of Medical Information & Library, Chinese Academy of Medical Sciences, 3rd Yabao Road, Chaoyang District, Beijing 100020, China.

Email: dai.tao@imicams.ac.cn