MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-126

Comparing Cancer Information Needs for Consumers in the US and China

Zongcheng Ji[a], Yaoyun Zhang[a], Jun Xu[a], Xiaoling Chen[a], Yonghui Wu[a], Hua Xu[a]

[a] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, USA

Abstract

Due to the differences in environments and cultures, consumers seeking cancer information in various regions of the world may have diverse needs. This study compares the cancer information needs for consumers in the US and China. Specifically, we first collected 1,000 cancer-related questions from Yahoo! Answers and Baidu Zhidao, respectively. Then, we developed a taxonomy of health information needs and manually classified the questions using the taxonomy. Finally, we analyzed the characteristics of information needs from consumers in both countries and summarized the differences between them. Our study demonstrated that although there are some common needs between consumers in the US and China, there are several significant differences between the two countries: the Chinese consumers are more likely to seek diagnosis and treatment online, while the US consumers prefer to seek common medical knowledge online.

Keywords:

Cancer; Information Needs; Taxonomy; Consumer Health Informatics

Introduction

Cancer is one of the leading causes of death worldwide. Tremendous efforts from the government, health providers, and pharmaceutical companies have been devoted to developing effective cancer therapies. With the initiative of precision medicine, a recent trend in cancer treatment is to provide personalized cancer therapy [1; 2], in which the proactive engagement from patients regarding treatment preference and healthcare data sharing plays a critical role. Therefore, multiple healthcare organizations (e.g., Mayo Clinic [1]) and government institutes (e.g., National Cancer Institute of US [2]) provide comprehensive online cancer knowledge services and have developed patient portals to answer healthcare questions and collect feedbacks, for better patient education and engagement in personalized cancer therapy.

Meanwhile, consumers (e.g., patients, family members, friends) seeking cancer information all over the world are also eager to know different cancer-related information, ranging from common knowledge about the causes, symptoms, and treatments to the most suitable healthcare providers and emerging new treatment options in the era of precision medicine. Moreover, people in different regions are attempting to find cancer knowledge and therapies worldwide to leverage better healthcare resources. Therefore, it is necessary to build automated informatics tools such as search engine or question answering systems that can integrate the rich resources and provide more efficient and effective cancer knowledge services for consumers. One essential step toward this goal is to first understand the specific cancer information needs of consumers, in order to provide the most helpful information.

Fortunately, such information needs are expressed in various social communities, such as online question answering communities where people can ask their questions and answer others' questions (e.g., Yahoo! Answers[3] in the US and Baidu Zhidao [4] in China). As time passed, these websites have accumulated a large amount of cancer-related questions and answers, which provide us invaluable data resources for understanding consumer health information needs. However, people from different regions with different cancer distributions, therapy development and cultures may have different information needs [3; 4]. For example, our preliminary study reveals that consumers in the US prefer to know the causes or risk factors of cancers, while consumers in China pay more attention to seek possible treatments online. Identifying the differences in information needs among consumers in different regions will facilitate the building of more advanced automated cancer knowledge services (e.g., search engines / question answering systems) which can provide knowledge / answers according to consumer-centered information needs. Furthermore, being aware of the different information needs required by the consumers will also facilitate cross-region collaborations for cancer diagnosis and treatment, such as the rapidly emerging international telemedicine.

There are several studies [5-11] that have analyzed cancer-related questions to understand consumer information needs. These studies mainly focused on very specific topics, including the information needs of post-treatment cancer patients [7; 9], patients in self-characterized illness phase [8], breast cancer patients [10], rare cancer patients [6], melanoma patients [11] and cancer patients' questions about pain [5]. However, no comprehensive study of consumer information needs of cancers has been conducted. Furthermore, the differences in information needs among comsumers in different regions of the world have not been previously investigated. This study compares the cancer information needs for consumers in different regions. As a starting point, we set the comparison between the US and China, because

[1] http://www.mayoclinic.org/
[2] https://www.cancer.gov/
[3] https://answers.yahoo.com/
[4] https://zhidao.baidu.com/

Table 1 – The detailed definitions for the 10 main categories of the taxonomy for consumer information needs related to cancers

Main Category	Description
Common Knowledge	Common medical knowledge, such as the definition, prevalence, etiology, prognosis or information resources of some conditions, etc.
Diagnosis	What are the possible conditions given certain symptoms, what are the possible symptoms given certain conditions, and knowledge about the tests such as the skin test, imaging and biopsy, etc.
Treatment	Treatments of diseases, which include the drug therapy, surgery, radiotherapy and targeted therapy, etc.
Prevention	The prevention of some cancers with drug, food or other methods.
Healthy Lifestyle	Healthy diet, mood control or other issues about how to keep healthy or help for recovery in daily life.
Health Provider Choosing	Information about choosing hospitals, departments, doctors or other healthcare organizations such as the support groups, associations and foundations, etc.
Second Opinion	Seek second opinions or suggestions from others.
Similar Experience Finding	Find someone who has similar experiences.
Finance	Information about financial support, such as the insurance, claim and charity, etc.
Other	Other questions that cannot fall into any categories above.

both countries have large scales of patients with different cancer distributions and cultures [3; 4]. Specifically, we first collected 1,000 cancer-related questions from Yahoo! Answers and Baidu Zhidao, respectively. Then, we developed a taxonomy of health information needs and manually classified the questions using the taxonomy. Finally, we analyzed the characteristics of information needs from consumers in both countries and summarized the differences between them. As far as we know, this is the first study to compare cancer information needs for consumers in the US and China. Our study will greatly benefit the development of automated cancer knowledge services for consumers as well as the cross-region collaboration for cancer therapy.

Methods

Datasets

We collected cancer-related questions from Yahoo! Answers and Baidu Zhidao, which represent the most popular online question answering communities in the US and China, respectively [12-16]. More specifically, we collected questions under the category of "Health / Diseases & Conditions / Cancer" from Yahoo! Answers and "医疗健康 / 肿瘤科 (Health Care / Oncology)" from Baidu Zhidao. To facilitate analysis, we took the *subject* field as the question without considering the *description* field, which contains the detailed description. Finally, we collected 9,043 and 166,469 questions in English and Chinese, respectively.

We randomly sampled 1,000 questions from each dataset to conduct the comparative study. As all the questions are user generated, some questions are not clearly described (e.g., Brain tumor?) or even not related to the cancer topic. We discarded these questions and randomly sampled new ones to keep the sample size at 1,000 for each dataset.

Taxonomy

To characterize consumer health information needs, a taxonomy is needed to classify the consumers' questions. The Taxonomy of Generic Clinical Questions (TGCQ) [17] is one of the most popular taxonomies for classifying health-related questions. Several studies [17; 18] show that TGCQ is useful for analyzing physicians' and case managers' information needs, but is not suitable for consumers' questions due to the difference in the information needs between the physicians and the consumers [19]. We are aware of one recent study [19] that developed a taxonomy mainly based on TGCQ for analyzing consumer health information needs on hypertension related questions. However, this taxonomy is incomplete or somewhat confusing for our case. For example, there is no category for the questions regarding seeking second opinions, finding similar experiences and asking financial support, where these three categories take up 11.7% of the English questions in our study. Furthermore, we also created a new category named *Common Knowledge*, which covers the definition, prevalence, etiology, prognosis or information resources of some conditions, etc. Finally, based on TGCQ and the taxonomy developed in the work of [19], we developed our new taxonomy, which consists of two levels of categories. The first level contains 10 main categories and the second level contains 28 subcategories. The detailed definitions for the 10 main categories are shown in *Table 1* and the 28 subcategories can be found in *Table 2*.

Annotation

We manually reviewed 100 questions from both the two datasets to develop an annotation guideline. Table 3 shows several examples from the annotation guideline. For instance, the question "Why does cancer treatment work for some and not others??" can be classified into the main category Treatment, but cannot fall into any other subcategories of Treatment (i.e., Drug Therapy, Surgery, Other Treatments and Treatment Seeking). Thus, we classify this question as Treatment→Other. Two annotators who are fluent in both English and Chinese manually classified the 2,000 questions (1,000 in English and 1,000 in Chinese) with the taxonomy we developed. The disagreements in the annotations were solved through group discussion including the two annotators.

Statistical Analysis

Cohen's kappa [20] was used to calculate inter-annotator agreement scores for the first level categories and the second level categories. We examined the frequency distribution of cancer-related questions among the first level categories and the second level categories for both the English dataset and the Chinese dataset. We also compared the frequency distribution across the English questions and the Chinese questions to identify the differences.

Results and Discussion

Taxonomy reliability

The two annotators manually annotated the 1,000 cancer-related questions from Yahoo! Answers (US) and 1,000 cancer-related questions from Baidu Zhidao (China) using the developed taxonomy. *Table 2* shows the frequency distribution over the two-level categories for the two datasets. The kappa measurement implemented in Stata was used to calculate the inter-annotator agreement between the two annotators.

Table 2 – Two-level categories of consumer information needs and their frequencies on caner-related questions

Main Category	Subcategory	Frequency on Yahoo Data		Frequency on Baidu Data	
Common Knowledge	*Definition*	423 (42.3%)	76 (7.6%)	244 (24.4%)	41 (4.1%)
	Prevalence		73 (7.3%)		31 (3.1%)
	Etiology		115 (11.5%)		86 (8.6%)
	Prognosis		64 (6.4%)		43 (4.3%)
	Information Seeking		47 (4.7%)		7 (0.7%)
	Other		48 (4.8%)		36 (3.6%)
Diagnosis	*Condition*	139 (13.9%)	56 (5.6%)	247 (24.7%)	131 (13.1%)
	Symptom		48 (4.8%)		33 (3.3%)
	Test		35 (3.5%)		83 (8.3%)
Treatment	*Drug Therapy*	212 (21.2%)	75 (7.5%)	352 (35.2%)	79 (7.9%)
	Surgery		26 (2.6%)		66 (6.6%)
	Other Therapy		27 (2.7%)		28 (2.8%)
	Treatment Seeking		62 (6.2%)		157 (15.7%)
	Other		22 (2.2%)		22 (2.2%)
Prevention	*Drug for prevention*	9 (0.9%)	1 (0.1%)	20 (2.0%)	1 (0.1%)
	Food for prevention		2 (0.2%)		6 (0.6%)
	Other		6 (0.6%)		13 (1.3%)
Healthy Lifestyle	*Diet*	35 (3.5%)	18 (1.8%)	71 (7.1%)	47 (4.7%)
	Mood Control		6 (0.6%)		5 (0.5%)
	Other		11 (1.1%)		19 (1.9%)
Health Provider Choosing	*Hospital*	32 (3.2%)	17 (1.7%)	48 (4.8%)	35 (3.5%)
	Department		0 (0.0%)		6 (0.6%)
	Doctor		7 (0.7%)		7 (0.7%)
	Other		8 (0.8%)		0 (0.0%)
Second Opinion		60 (6.0%)	60 (6.0%)	7 (0.7%)	7 (0.7%)
Similar Experience Finding		33 (3.3%)	33 (3.3%)	2 (0.2%)	2 (0.2%)
Finance		24 (2.4%)	24 (2.4%)	1 (0.1%)	1 (0.1%)
Other		33 (3.3%)	33 (3.3%)	8 (0.8%)	8 (0.8%)

Table 3 – Examples from the annotation guideline of consumer information needs related to cancers

Category	Guideline / Generic Types of Questions	Example Questions (English translations are given following the Chinese questions.)
Common Knowledge → Etiology	Causes such as genetic inheritance / risk factors of some conditions	Does salt lead to lung cancer? 血管瘤一定是先天性的吗? (Are all the cases of Hemangioma congenital?)
Treatment → Treatment seeking	Questions about seeking help or treatment for some conditions.	How do you treat breast cancer? 平滑肌肉瘤如何治疗 (How to treat leiomyosarcoma)
Treatment → Other	Questions about treatment, which cannot fall into any other subcategories of Treatment.	Why does cancer treatment work for some and not others?? 淋巴瘤可以根治吗 (Is lymphoma curable?)
Second Opinion	Questions of seeking second opinions or suggestions.	If you have just had a very close neighbor die of liver cancer..what do you do? 去看望胃癌病人，带什么比较好？（食物、水果） (To visit patients with gastric cancer, which is better to bring? Food or fruits?)
Similar Experience Finding	Questions of finding similar experiences.	Has anyone suffered from choriocarcinoma? 生命丝带对乳房腺瘤有用吗？用过的朋友请进 (Is Life Ribbon useful for breast adenoma? Those who have used it, please come here)
Finance	Questions about financial support, such as the insurance, claim and charity, etc.	I would like to know if i can get affordable insurance for melanoma treatment? 军人得癌病，需要终生服药，复员时有相应政策吗？ (Is there any demobilization policy for soldiers who have cancers and need life-long medication?)

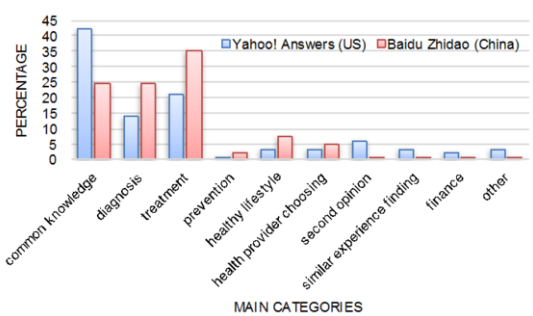

Figure 1 – Comparisons of frequency distributions over the 10 main categories for the questions from the US and China

For the first level categories, the two annotators achieved a Kappa score of 0.9105 and 0.9502 for the English questions and the Chinese questions, respectively. For the second level categories, the two annotators achieved a Kappa score of 0.8976 and 0.8798 for the English questions and Chinese questions, respectively. The good kappa scores indicated that the two annotators achieved high agreement on both of the two-level categories.

Frequency distribution of cancer information needs in US

Among the 1,000 English questions from Yahoo! Answers, the most frequent main category is *Common Knowledge* (42.3%), with the most frequent subcategory being *Etiology*, indicating that consumers in the US are more concerned with the cause or risk factors of diseases. The second most frequent main category is *Treatment* (21.2%), where the most frequent

subcategory is *Drug Therapy*. The third most frequent main category is *Diagnosis* (13.9%), where most of the questions are about the subcategory *Condition* (i.e., what are the possible conditions given the symptoms). For other main categories, the frequency distributions are 6% for *Second Opinion*, 3.5% for *Healthy Lifestyle*, 3.3% for *Similar Experience Finding*, 3.2% for *Healthcare Provider Choosing*, 2.4% for *Finance* and 0.9% for *Prevention*.

Frequency distribution of cancer information needs in China

Among the 1,000 Chinese questions from Baidu Zhidao, the most frequent main category is *Treatment* (35.2%), where the most frequent subcategory is *Treatment Seeking*, indicating that consumers in China care more about seeking possible treatments online. The second most frequent main category is *Diagnosis* (24.7%), where the most frequent subcategory is *Condition*. The third most frequent main category is *Common Knowledge* (24.4%), where the most frequent subcategory is *Etiology*. The frequency distribution for other main categories are 7.1% for *Healthy Lifestyle*, 4.8% for *Healthcare Provider Choosing* and 2% for *Prevention*. There are very few questions about requesting second opinions, finding similar experiences and asking for help with financial support.

Comparisons of cancer information needs between China and US

Figure 1 and *Figure 2* shows the comparisons of frequency distributions over the 10 main categories and the 28 subcategories for the two datasets, respectively. From the two figures, we can find that the most frequent main categories for both datasets are *Common Knowledge*, *Diagnosis* and *Treatment*, which covers 77.4% and 84.3% of the questions in the two datasets, respectively. This indicates that consumers in the US and China have the similar concerns even with different healthcare systems and cultures.

Among the top three most frequent main categories of both datasets, we found several differences between the US and China. For the US consumers, the most frequent category is *Common Knowledge*, while it is *Treatment* for the Chinese consumers. This may indicate that the US consumers are more likely to ask about common medical knowledge, whereas the Chinese consumers are more likely to ask about the possible treatments for specific diseases. Among the subcategories of the top three frequent main categories, the Chinese consumers are more likely to ask about the *Treatment Seeking and Condition* (i.e., what are the possible conditions given the symptoms); whereas the US consumers are more likely to ask about the common knowledge related to *Etiology*. Why the Chinese consumers are more likely to seek diagnosis and therapies through online communities is not certain, but it is interesting and worth conducting further investigation. Potential reasons may be the differences between the healthcare systems. For example, few Chinese have primary care providers, whereas many US patients can get useful information from their family doctors.

Beside the differences among the three common categories, we also found that the US consumers are more likely to ask questions about *Second Opinion*, *Similar Experience Finding* and *Finance* than the Chinese consumers (11.7% vs. 1%). This indicates that the US consumers are prone to exchange personal opinions and seek financial support, while the Chinese consumers not. This might be related to the differences between health insurance systems. Most of the Chinese consumers have government sponsored insurance, whereas the US patients are involved in very complex insurance systems. Again, further investigation should be taken to determine why this difference exists. On the other side, the Chinese consumers care more about the *Diet* for *Healthy Lifestyle* (4.7%), which indicates that more consumers in China consider the diet to be helpful for cancer treatment. This also reflects one of the Chinese traditional cultures – many Chinese consumers believe in food therapy.

Limitations

This study has several limitations. First, we limited the sources of English questions to Yahoo! Answers and Chinese questions to Baidu Zhidao. These two sources have different usability designs and demographic characteristics, which should be taken into account in our future work. In addition, questions from other online question answering communities will also be collected and analyzed in the next step. Second, we only focused on the general cancer domain and did not differentiate between different cancer subtypes. Our future work includes developing automated methods to mine information needs for different cancer subtypes. We are also planning to integrate this work into a medical question answering system to utilize the information needs embedded in the questions.

Conclusions

In this study, we identified the cancer information needs for consumers both in the US and China, and compared the differences. We developed a taxonomy of consumer health information needs regarding cancer and annotated two cancer-related datasets with the taxonomy. Based on the quantitative analysis, we reported some interesting observations. To the best of our knowledge, this is the first study focused on consumer information needs for cancer between the US and China. Our study will benefit the development of automated cancer-related search engine / question answering systems for consumers as well as the cross-region collaboration for cancer therapy.

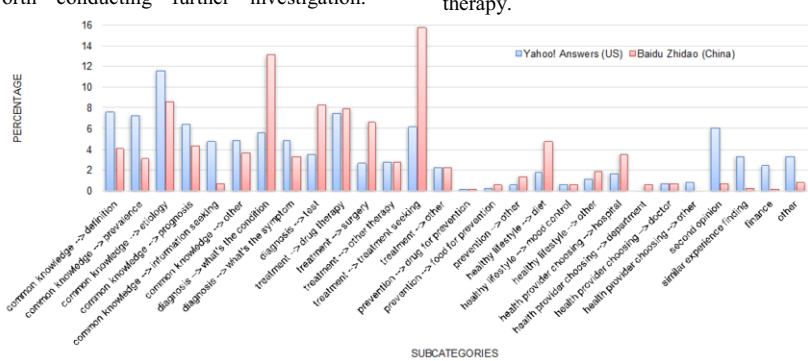

Figure 2 – Comparisons of frequency distributions over the 28 subcategories for the questions from the US and China

Acknowledgements

This study is supported in part by grants from NLM 2R01LM010681-05, NIGMS 1R01GM103859, and 1R01GM102282.

References

[1] S. Heath, Facilitating Patient Participation in Precision Medicine, in: https://goo.gl/dKrXDC, 2016.
[2] C. Petersen, The Future of Patient Engagement in the Governance of Shared Data, *eGEMs* **4** (2016).
[3] W. Chen, R. Zheng, P.D. Baade, S. Zhang, H. Zeng, F. Bray, A. Jemal, X.Q. Yu, and J. He, Cancer statistics in China, 2015, *CA: a cancer journal for clinicians* **66** (2016), 115-132.
[4] R.L. Siegel, K.D. Miller, and A. Jemal, Cancer statistics, 2015, *CA: a cancer journal for clinicians* **65** (2015), 5-29.
[5] J.L. Bender, J. Hohenadel, J. Wong, J. Katz, L.E. Ferris, C. Shobbrook, D. Warr, and A.R. Jadad, What patients with cancer want to know about pain: a qualitative study, *Journal of pain and symptom management* **35** (2008), 177-187.
[6] R. Falotico, C. Liberati, and P. Zappa, Identifying Oncological Patient Information Needs to Improve e‐Health Communication: a preliminary text‐mining analysis, *Quality and Reliability Engineering International* **31** (2015), 1115-1126.
[7] E.M. Galarce, S. Ramanadhan, J. Weeks, E.C. Schneider, S.W. Gray, and K. Viswanath, Class, race, ethnicity and information needs in post-treatment cancer patients, *Patient education and counseling* **85** (2011), 432-439.
[8] J.E.Z.D.W. Pratt, Self-characterized illness phase and information needs of participants in an online cancer forum, (2015).
[9] M. Shea–Budgell, X. Kostaras, K. Myhill, and N. Hagen, Information needs and sources of information for patients during cancer follow-up, *Current Oncology* **21** (2014), 165.
[10] Y. Zhang and H. Xu, Understanding consumers ' information needs about breast cancer by analyzing online questions, in: *AMIA*, 2013, p. 1578.
[11] A. Molassiotis, L. Brunton, J. Hodgetts, A.C. Green, V.L. Beesley, C. Mulatero, J.A. Newton-Bishop, and P. Lorigan, Prevalence and correlates of unmet supportive care needs in patients with resected invasive cutaneous melanoma, *Annals of Oncology* **25** (2014), 2052-2058.
[12] D. Wu and D. He, A study on Q&A services between community-based question answering and collaborative digital reference in two languages, *IAENG International Journal of Computer Science* **40** (2013), 110-116.
[13] Z. Ji, F. Xu, and B. Wang, A category-integrated language model for question retrieval in community question answering, in: *AIRS*, 2012, pp. 14-25.
[14] F. Xu, Z. Ji, and B. Wang, Dual role model for question recommendation in community question answering, in: *SIGIR*, 2012, pp. 771-780.
[15] Z. Ji and B. Wang, Learning to rank for question routing in community question answering, in: *CIKM*, 2013, pp. 2363-2368.
[16] Z. Ji, F. Xu, B. Wang, and B. He, Question-answer topic model for question retrieval in community question answering, in: *CIKM*, 2012, pp. 2471-2474.
[17] J.W. Ely, J.A. Osheroff, P.N. Gorman, M.H. Ebell, M.L. Chambliss, E.A. Pifer, and P.Z. Stavri, A taxonomy of generic clinical questions: classification study, *BMJ* **321** (2000), 429-432.
[18] R. Schnall, J.J. Cimino, L.M. Currie, and S. Bakken, Information needs of case managers caring for persons living with HIV, *JAMIA* **18** (2011), 305-308.
[19] H. Guo, J. Li, and T. Dai, Consumer Health Information Needs and Question Classification: Analysis of Hypertension Related Questions Asked by Consumers on a Chinese Health Website, in: *MedInfo*, 2015, pp. 810-814.
[20] J. Cohen, A Coefficient of Agreement for Nominal Scales, *Educational and psychological measurement* **20** (1960), 37-46.

Address for correspondence

Hua Xu

Hua.Xu@uth.tmc.edu