Interactive use of online health resources: a comparison of consumer and professional questions

RECEIVED 25 September 2015
REVISED 11 November 2015
ACCEPTED 30 November 2015
PUBLISHED ONLINE FIRST 4 May 2016

Kirk Roberts and Dina Demner-Fushman

ABSTRACT

Objective To understand how consumer questions on online resources differ from questions asked by professionals, and how such consumer questions differ across resources.

Materials and Methods Ten online question corpora, 5 consumer and 5 professional, with a combined total of over 40 000 questions, were analyzed using a variety of natural language processing techniques. These techniques analyze questions at the lexical, syntactic, and semantic levels, exposing differences in both form and content.

Results Consumer questions tend to be longer than professional questions, more closely resemble open-domain language, and focus far more on medical problems. Consumers ask more sub-questions, provide far more background information, and ask different types of questions than professionals. Furthermore, there is substantial variance of these factors between the different consumer corpora.

Discussion The form of consumer questions is highly dependent upon the individual online resource, especially in the amount of background information provided. Professionals, on the other hand, provide very little background information and often ask much shorter questions. The content of consumer questions is also highly dependent upon the resource. While professional questions commonly discuss treatments and tests, consumer questions focus disproportionately on symptoms and diseases. Further, consumers place far more emphasis on certain types of health problems (eg, sexual health).

Conclusion Websites for consumers to submit health questions are a popular online resource filling important gaps in consumer health information. By analyzing how consumers write questions on these resources, we can better understand these gaps and create solutions for improving information access.

This article is part of the Special Focus on Person-Generated Health and Wellness Data, which published in the May 2016 issue, Volume 23, Issue 3.

Keywords: consumer health informatics, online information seeking, consumer language, question answering

RESEARCH AND APPLICATIONS

BACKGROUND AND SIGNIFICANCE

Patients and caregivers (health information consumers) are increasingly more active in decision-making about their own and their family members' health.[1] This involvement often motivates consumers to ask questions and seek information online. Consumers' information needs and access to health information, and the role and coverage of the health-related resources, are partially captured in consumer interactions with online resources. These interactions involve the majority of US adults: among the 87% of US adults who use the Internet, 72% look online for health information, the majority starting with a search engine query.[2] Clinicians and other health professionals, the primary resource for consumers' health information, also frequently turn to online resources when looking for answers to their questions.[3]

Insights into consumer and professional information needs and coverage of health issues provided by online resources are most frequently gleaned from search engine log analyses and surveys. In a study of relationships between online health-seeking behaviors and in-world health care utilization, White and Horvitz[4] analyzed data from surveys and online search logs and found that information needs change from exploration of diseases and symptoms and searches for information about doctors and facilities prior to a visit to a health facility to searches for treatments and benign symptoms after the visit. They also report differences in search behavior based on the lower and higher levels of domain knowledge (analogous to the difference between consumers and professionals). There are also notable terminological differences between consumers and professionals, which has led to the development of terminologies specifically for consumers,[5] although some domains have more overlap than others.[6]

Not all questions have easily accessible answers: in a 2004 survey, 97 subjects found answers to 30% of their questions and partial answers to another 33% in MedlinePlus and other related websites.[7] Furthermore, although logs are relatively available and useful sources, they present user search strategies and general information needs but not specific needs. Alternatively, questions posted to online forums or question answering (QA) sites not only reflect the information needs more fully and explicitly, but also are commonly the next step after failing to find information via search engines.[8]

Many studies have analyzed a small number of consumer questions, often further limiting the analysis to specific topics and convenience samples. White[9] analyzed 365 mailing list questions for type and subject. Oh et al.[10] studied 72 community QA questions for linguistic style and sentiment. Zhang[11] manually categorized 276 community QA questions for motivation, temporality, and cognitive representation. Slaughter et al.[12] manually annotated semantic relations on 12 consumer questions with professional answers. In contrast to these studies, we were able to analyze tens of thousands of consumer and professional

Correspondence to Kirk Roberts, Lister Hill National Center for Biomedical Communications, US National Library of Medicine, National Institutes of Health 8600 Rockville Pike, Building 38A/1003H Bethesda, MD, 20894, USA; kirk.roberts@nih.gov

questions using natural language processing (NLP). This is the first large-scale analysis of consumer and professional health questions. While this scaled analysis is not able to recognize high-level characteristics such as style and motivation, it is able to remove much of the sampling bias of smaller studies by studying both consumers and professionals across a wide variety of online resources.

Further, automatic understanding of these questions could improve the consumer online experience by retrieving answers best matching the fine-grained needs.[13] Automatic question understanding methods could themselves be improved by discerning the nature and characteristics of online health questions.

This article helps to elicit the characteristics of online health-seeking behavior in the form of questions. We use a variety of NLP techniques, including several novel methods developed by the authors, to analyze 10 question corpora containing approximately 30 000 consumer and 10 000 professional questions. Specifically, we seek to qualify how:

a. Consumer questions in general differ from professional questions in form and content, and
b. Consumer questions differ across resources (eg, community QA, forums, emails) in form and content.

We are not aware of any such consumer question study approaching the size of that presented here in terms of (i) total number of questions, (ii) number and variety of corpora, and (iii) depth and breadth of NLP techniques. Note that our goal is not to use NLP to automatically classify consumer versus professional questions, as has been done by others,[14] since consumers and professionals often naturally gravitate to different resources. Instead, our goal is to use NLP to analyze far more data than could possibly be done manually and identify the linguistic tendencies of consumers as a whole and within specific resources. The question analysis approach in this work provides a means for designing and evaluating QA systems.

MATERIALS AND METHODS

In order to compare consumer and professional questions, we seek to compare questions using various linguistic levels (lexical, syntactic, semantic, etc.) across a wide variety of online resources. The Corpora subsection below describes these resources, while the Analysis Methods subsection describes the levels of linguistic analysis useful for comparing health questions.

Corpora

We identified 5 types of resources: community QA (consumers answer publicly, and the best answer is ranked/selected); curated QA (professionals selectively answer publicly); forum (consumers answer publicly in a conversation); email (professionals answer privately); and point-of-care (stream-of-conscience clinical questions without a precise audience).

Five health consumer corpora were obtained: (1) Yahoo! Answers (**YANS**), a popular community QA website where questions are posed and answered by consumers; 4.5 million questions, with answers, are publicly available for academic research purposes (http://webscope.sandbox.yahoo.com/).[15] The dataset contains 49 582 questions under the Diseases & Conditions subcategory (under Health), of which we randomly sampled 10 000. (2) WebMD Community (**WEBC**), a consumer health forum hosted by WebMD (http://exchanges.webmd.com/). Consumers post questions on the forum, resulting in conversations that differ in style from the community QA sites. Over 230 000 forum posts were downloaded from 209 subforums. Since the number of topics in each subforum is skewed toward parenting and pregnancy, we performed a stratified sampling of each subforum to obtain 10 000 questions that reflect the breadth of topics. (3) Doctorspring (**DSPR**), a curated QA website where consumers submit questions to be answered by a health professional, for a fee (http://www.doctorspring.com/questions-home). We downloaded 811 questions from the website. (4) Genetic and Rare Diseases Information Center (**GARD**), a curated QA website where consumers submit questions to be answered by NIH staff (http://rarediseases.info.nih.gov/). We obtained 1467 questions from GARD. (5) NLM Consumer Health Questions (**NLMC**), which contains questions about diseases, conditions, and therapies submitted to NLM's websites or via e-mail. The question submitters self-identify as "General Public." Answers are provided by NLM staff via e-mail. We obtained 7164 consumer questions submitted to NLM between 2010 and 2014.

Five health professional corpora were obtained: (1) Parkhurst Exchange (**PHST**), a journal for physicians that maintains a curated QA resource (http://www.parkhurstexchange.com). We downloaded 5290 questions from the website. (2) Journal of Family Practice (**JFPQ**), another journal with curated questions targeted toward specific cases (http://www.jfponline.com/articles/clinical-inquiries.html). We downloaded 601 questions from the website. (3) Clinical Questions (**CLIQ**), collected by Ely et al.[16,17] and D'Alessandro et al.[18] at the point of care, either during direct observation or by phone interview. There are 4654 questions in the collection (http://clinques.nlm.nih.gov). (4) Questions posed during an evaluation of PubMed on Tap (**PMOT**), which provides point-of-care access to PubMed using handheld devices.[19] These questions more closely resemble keyword queries, though many are well-formed questions. We obtained 521 questions from this collection. (5) NLM Professional Health Questions (**NLMP**), similar to NLMC, but for users who self-identify as a "Health Professional" or "Researcher/Scientist." We obtained 740 professional questions submitted to NLM between 2010 and 2014.

Appendix A in the supplemental data contains examples from these 10 corpora. The corpora were chosen based on our awareness of their availability. We were unable to find a suitable general community QA/forum website for professionals, or point-of-care questions for consumers.

Analysis methods

The questions in each corpus were analyzed using a battery of techniques designed to represent various types of lexical, syntactic, and semantic information. The techniques are summarized below. Implementation details are provided in Appendix B in the online supplemental data.

Lexical

Question length was measured in words, tokens, and sentences. Word length was measured in characters. Sentence length was measured in tokens. Finally, the number of capitalized words (first character only) was measured.

Readability

Three metrics were applied: (a) Gunning fog index,[20] (b) Flesch reading ease,[21] and (c) Flesch-Kincaid grade level.[22] These metrics rely on statistics such as sentence length, word length, and word complexity. Additionally, the number of misspelled words was estimated using several large corpora.

Language Model

Questions were evaluated with 2 trigram language models.[23] The first model is an open-domain language model built from newswire[24] and

Figure 1: Example bipolar questions written by (a) a consumer and (b) a professional.

(a) **WEBC Question 18:**
hello fellow forumers. I have been diagnosed with bipoar 1 with psychotic features since 18 years old. (1 am 32 now) i am having a dysphoric hypomania right now according to my psychiatrist)- it is under control and getting better....but anyway, how can i have a HYPO mania if i am bipolar 1? ..and...doesnt everyone with bipolar 1 have psychotic features? why do they have to add that to the end of my diagnosis? isn't dysphoric hypomia the same thing as a mixed episode? ive suposedly had these before.....So do you think my diagnosis of bipolar 1 with psychotic features is correct? It might help to know my history- yes i have psychosis, hard to admit. many of the paranoia and delusions i still besieve are real, no matter what people tell me., I also (multiple times) have been the manic,and depressed,had mixed,episodes, this: (dysphoric hypomania) (which would change my diagnosis, right?

(b) **PHST Question 536:**
Are antidepressants contraindicated in bipolar illness even when combined with mood-stabilizing medication?

Wikipedia. Due to the large size of both corpora, a 10% sample of the sentences was used to build the model. The second model is a medical language model built from a 20% sample of PubMed Central (http://www.ncbi.nlm.nih.gov/pmc/). Both the document-level and sentence-level log probabilities were measured.

Semantic Types

Both open-domain and medical semantic types were extracted. For open-domain types, a named entity recognizer extracted PERSON, ORGANIZATION, LOCATION, NUMERIC, TIMEDATE, and MISC types. For medical types, a dictionary lookup was performed using the Unified Medical Language System (UMLS) Metathesaurus.[25] To highlight certain facets of medical language, 2 views of UMLS were employed: (a) semantic types grouped into PROBLEM, TREATMENT, and TEST and (b) individual terminologies in MeSH (http://www.nlm.nih.gov/mesh/), SNOMED-CT,[26] and the Consumer Health Vocabulary.[5]

Question Decomposition

Many questions were paragraphs containing several subquestions. For instance, Figure 1 (a) shows a WEBC question containing at least 6 subquestions. To recognize the number of subquestions and background sentences, the system described in Roberts et al.[27] was used to syntactically decompose the questions. Next, each subquestion was classified into 1 of 13 types.[28] The common question types include INFORMATION (general information), MANAGEMENT (treatment and prevention), and SUSCEPTIBILITY (how a disease is acquired or who is vulnerable). For more details on the question types, including how they were created, see Roberts et al.[29] Finally, we counted the questions that started with typical wh-word question stems (*who*, *what*, *when*, *where*, *why*, and *how*) to measure the question's surface-level type.

Topics

Topic modeling techniques can provide a useful summary of large amounts of unstructured text. We utilized Latent Dirichlet Allocation (LDA)[30] with 10 topics in order to compare the subject matter across corpora. Separate topic models were built using question words and UMLS terms.

Classification

Finally, to assess the relative importance of these metrics as discriminators between consumer and professional questions, we created a logistic regression model using the metrics described above as features. Again, unlike Liu et al.,[14] our goal was not to create the best possible classifier, but rather to determine the relative importance of the analysis methods as an empirical indicator of how consumer and professional questions differ.

RESULTS

The results of the analyses are shown in Tables 1 and 2.

Lexical

Consumers tend to ask longer questions (in the range of 37–106 tokens for consumers versus 11–62 for professionals), though this appears to be primarily an effect of the resource. Among similar resources, the divide was smaller: QA websites had some variation (37–100 versus 11–36), while the NLM questions varied little (70 versus 62). On the other hand, point-of-care questions, for which we have no comparable consumer corpus, were quite short (11–24), and forum questions, for which we have no comparable professional corpus, were quite long (106). Similar effects can be seen with sentences, where most professional questions had 1 or 2 sentences (except NLMP), while consumer questions tended to have 3 or more (2.8–6.9). Word length was shorter for consumers (4.0–4.7 characters versus 4.5–5.5), suggesting a less developed vocabulary, perhaps resulting in more words to describe an information need. There are few discernible differences between how consumers and professionals capitalize words (9.1–14.4 versus 8.8–14.4).

Readability

According to the metrics, consumer questions are more readable than professional questions. The fog index for consumers is lower (9.0–11.9 versus 12.2–14.8), as is the grade level (6.2–9.2 versus 9.2–11.9), implying less education is needed to comprehend the questions. Similarly, the reading ease is higher for consumers (49.6–75.4 versus 32.3–49.9). WEBC is consistently the most readable consumer corpus, and GARD is consistently the least readable. Thus, consumer-to-consumer websites tend to be the most readable, while professional-to-professional medical journals the least. Rates of misspelling are higher in consumer questions than professional questions, except for NLMP (see Discussion section).

Table 1. Lexical, Readability, Language Model, and Semantic Type statistics for the 10 corpora

	YANS	WEBC	DSPR	GARD	NLMC	PHST	JFPQ	CLIQ	PMOT	NLMP
Type	Comm. QA	Forum	Curated QA	Curated QA	Email	Curated QA	Curated QA	Point of Care	Point of Care	Email
# Questions	10 000	10 000	811	1467	7164	5290	601	4654	521	740
Avg Question Length (Tokens)	43 ($\sigma = 46$)	106 (57)	100 (77)	37 (24)	70 (76)	36 (22)	11 (3)	24 (16)	11 (4)	62 (74)
Avg Question Length (Words)	37 (41)	92 (50)	89 (69)	33 (22)	60 (65)	30 (19)	10 (3)	20 (13)	10 (4)	53 (62)
Avg Question Length (Sentences)	3.1 (2.9)	6.9 (3.9)	6.8 (5.2)	2.8 (1.7)	4.7 (5.0)	2.0 (1.2)	1.0 (0.0)	1.7 (1.1)	1.1 (0.4)	4.3 (4.4)
Avg Word Length (Characters)	4.1 (2.8)	4.0 (2.2)	4.1 (2.3)	4.7 (2.8)	4.2 (2.6)	4.9 (3.0)	5.5 (3.0)	4.7 (5.3)	5.3 (3.2)	4.5 (2.7)
Avg Sentence L ength (Tokens)	14 (11)	15 (11)	15 (10)	13 (6)	15 (16)	18 (9)	11 (3)	14 (7)	10 (4)	14 (16)
Avg % Capitalized	9.1 (10.6)	11.7 (6.3)	12.2 (5.3)	14.4 (7.4)	12.4 (11.4)	8.8 (4.5)	11.9 (4.9)	13.9 (9.2)	14.4 (7.4)	12.7 (13.9)
Avg Fog Index	9.2 (4.8)	9.0 (3.8)	10.2 (4.2)	11.9 (4.1)	11.0 (5.6)	14.8 (4.4)	14.4 (5.7)	12.2 (5.0)	13.5 (6.0)	12.4 (5.7)
Avg Reading Ease	69.8 (51.5)	75.4 (15.3)	70.2 (14.7)	49.6 (25.9)	61.6 (27.4)	37.0 (22.9)	32.3 (33.0)	49.3 (28.1)	33.1 (39.5)	49.9 (30.3)
Avg Grade Level	6.3 (7.3)	6.2 (3.6)	7.1 (3.9)	9.2 (3.7)	7.9 (5.1)	11.9 (3.8)	11.2 (4.5)	9.2 (4.3)	10.8 (5.6)	9.3 (4.8)
Avg % Misspellings	1.4 (4.8)	0.6 (1.8)	0.4 (1.2)	0.1 (0.5)	1.9 (4.5)	0.1 (0.7)	0.0 (0.9)	0.1 (0.8)	1.1 (0.4)	2.1 (4.0)
Avg Open Log-Prob (Document)	−128.2	−309.7	−283.1	−99.9	−218.7	−108.3	−39.6	−71.8	−38.7	−201.6
Avg Open Log-Prob (Sentence)	−41.3	−44.3	−41.6	−35.6	−46.6	−54.6	−39.5	−40.8	−34.4	−46.8
Avg Medical Log-Prob (Document)	−136.3	−335.2	−308.1	−100.7	−222.7	−95.1	−33.5	−67.5	−35.5	−197.4
Avg Medical Log-Prob (Sentence)	−43.9	−48.3	−45.3	−35.9	−47.4	−47.9	−33.5	−38.3	−31.5	−45.8
Avg % Person	0.2 (1.8)	0.3 (1.0)	0.1 (0.4)	0.8 (2.6)	0.5 (2.2)	0.2 (1.4)	0.1 (1.6)	0.4 (2.1)	0.4 (2.3)	0.6 (1.9)
Avg % Organization	0.1 (1.3)	0.1 (0.8)	0.0 (0.3)	0.1 (0.8)	0.2 (1.4)	0.1 (0.9)	0.1 (0.7)	0.2 (1.8)	0.1 (1.0)	0.3 (1.2)
Avg % Location	0.2 (1.6)	0.1 (0.6)	0.0 (0.2)	0.2 (1.6)	0.3 (1.3)	0.1 (0.8)	0.0 (0.0)	0.2 (1.2)	0.1 (1.1)	0.3 (1.2)
Avg % Numeric	1.2 (3.9)	1.8 (3.2)	1.1 (2.5)	0.7 (1.9)	2.5 (4.7)	1.1 (2.9)	0.2 (1.4)	0.9 (2.9)	0.3 (1.9)	2.2 (4.5)
Avg % Timedate	1.7 (3.9)	4.1 (4.4)	4.5 (4.2)	1.9 (3.9)	2.9 (4.6)	2.4 (4.8)	0.3 (2.2)	3.7 (7.5)	0.4 (1.8)	2.3 (4.5)
Avg % Misc	0.2 (1.6)	0.1 (0.8)	0.1 (0.6)	0.5 (2.6)	0.2 (0.9)	0.2 (1.4)	0.2 (2.0)	0.4 (2.3)	0.3 (1.7)	0.4 (2.9)

Numbers in parentheses are standard deviations. Percentages are out of 100.

Language Model

Language model probabilities are highly dependent on text length, making it infeasible to make cross-corpora inferences. However, running 2 language models on the same corpus allows us to infer which training corpus (newswire + Wikipedia versus PubMed Central) the question corpus more closely resembles. The models estimate the probability of a text given the training corpus using a simplified n-gram assumption. Since the probability of any given text is quite small, we provide the log-probability. Smaller negative numbers are thus more likely than larger negative numbers. In these experiments, every consumer corpus was judged more probable by the open-domain model, and every professional corpus was judged more probable by the medical model.

Semantic Types

Open-domain named entities (proper names) appear to be relatively rare in health questions for both consumers and professionals. The use of times and dates, though, is quite common, especially for consumers (1.7–4.5% versus 0.3–3.7%). Temporal expressions are often used to build a disease narrative (eg, *5 weeks ago*) or describe symptoms (*every few hours*) and patient characteristics (*27-year-old*). The 2 corpora with the greatest concentration of Timedate entities, WEBC (4.1% of tokens) and DSPR (4.5%), have the longest questions, suggesting that consumers use the additional space to add a more detailed temporal narrative.

Medical semantic types are more common than the open-domain entities. Problems are the most frequent type, then Treatments, with Tests being least frequent. The consumer corpora have an average Problem:Treatment:Test ratio of approximately 16:5:1, whereas the professional corpora have an average ratio of 6:4:1. From this, we can see that consumers use appreciably more space discussing problems and rarely discuss tests. Another means of comparing the medical language is to use the constituent terminologies in UMLS, specifically MeSH (intended for scientific articles), SNOMED-CT (professionals), and the Consumer Health Vocabulary (CHV, consumers). MeSH is smaller and probably more neutral. Regardless of vocabulary,

Table 2. Semantic Type and Question Decomposition statistics for the 10 corpora

	YANS	WEBC	DSPR	GARD	NLMC	PHST	JFPQ	CLIQ	PMOT	NLMP
Avg % UMLS Problems	8.5 ($\sigma = 8.8$)	4.2 (3.4)	6.5 (4.9)	12.8 (9.4)	7.8 (9.2)	8.3 (7.3)	14.2 (12.2)	9.2 (9.8)	11.5 (11.8)	8.4 (9.2)
Avg % UMLS Treatments	3.2 (5.5)	2.2 (2.5)	2.1 (2.9)	2.4 (3.8)	3.4 (5.2)	6.1 (6.3)	8.1 (9.4)	5.6 (7.4)	6.5 (8.7)	4.3 (6.6)
Avg % UMLS Tests	0.5 (2.6)	0.4 (1.1)	0.4 (1.4)	0.6 (2.1)	0.6 (5.5)	2.0 (4.1)	1.9 (6.0)	2.2 (5.4)	1.7 (5.6)	0.9 (3.3)
Avg % MeSH	12.6 (9.5)	7.0 (4.3)	9.7 (4.7)	15.9 (9.2)	12.8 (10.1)	16.6 (8.1)	24.8 (12.0)	17.7 (10.1)	20.7 (13.5)	15.2 (12.2)
Avg % SNOMED	20.6 (10.7)	16.6 (6.2)	21.1 (6.0)	23.2 (9.3)	22.0 (11.2)	27.3 (9.3)	29.0 (13.6)	27.1 (11.0)	27.5 (14.7)	23.9 (12.3)
Avg % CHV	26.4 (11.4)	21.1 (6.3)	26.2 (6.2)	28.9 (9.7)	27.2 (12.4)	34.1 (9.0)	41.8 (12.8)	33.4 (11.0)	35.7 (16.0)	30.4 (14.2)
Avg # Subquestions	1.7 (1.1)	2.4 (1.7)	2.0 (1.5)	1.7 (1.0)	1.8 (1.8)	1.4 (0.7)	1.0 (0.2)	1.2 (0.6)	1.1 (0.4)	1.7 (1.3)
Avg # Background Sent.	1.3 (2.2)	4.2 (3.1)	4.5 (4.2)	1.2 (1.3)	2.4 (3.4)	0.7 (1.0)	0.0 (0.0)	0.6 (1.0)	0.1 (0.3)	2.2 (3.4)
Avg # Ignore Sent.	0.2 (0.5)	0.4 (0.9)	0.4 (0.7)	0.0 (0.2)	0.5 (1.3)	0.0 (0.1)	0.0 (0.0)	0.0 (0.1)	0.0 (0.0)	0.5 (1.2)
Avg % Anatomy	0.1 (3.0)	0.1 (1.7)	0.0 (1.2)	0.2 (4.0)	0.2 (4.3)	0.0 (2.6)	0.0 (0.0)	0.1 (3.3)	0.0 (0.0)	0.0 (0.0)
Avg % Cause	4.9 (19.3)	3.1 (14.0)	4.0 (15.8)	4.0 (17.1)	3.0 (14.4)	3.9 (17.4)	1.8 (13.4)	7.0 (24.2)	9.1 (27.9)	3.0 (13.8)
Avg % Complication	0.3 (4.1)	0.2 (3.3)	0.3 (4.4)	1.4 (10.7)	0.3 (4.3)	1.2 (10.0)	1.7 (12.8)	0.5 (6.7)	0.4 (6.3)	0.4 (5.1)
Avg % Diagnosis	4.8 (19.2)	4.2 (15.7)	5.1 (17.2)	7.5 (23.1)	4.4 (17.0)	11.0 (28.9)	11.8 (32.1)	9.4 (27.6)	7.4 (25.7)	7.8 (24.0)
Avg % Information	17.3 (32.6)	17.7 (29.3)	10.4 (22.7)	19.2 (35.7)	13.1 (27.8)	6.1 (21.4)	4.4 (20.0)	15.7 (34.3)	14.0 (33.6)	14.7 (29.3)
Avg % Management	30.6 (40.8)	26.9 (34.5)	31.8 (37.2)	23.9 (38.1)	31.1 (38.7)	43.2 (45.4)	57.5 (49.0)	34.4 (45.2)	34.1 (46.3)	33.1 (39.5)
Avg % Manifestation	2.9 (14.9)	1.3 (8.8)	3.3 (15.0)	2.7 (13.8)	1.7 (11.0)	1.3 (10.4)	1.2 (10.7)	1.3 (10.7)	0.9 (8.6)	1.1 (9.9)
Avg % NotDisease	0.2 (3.4)	0.4 (5.1)	0.3 (3.6)	0.5 (6.8)	0.2 (3.7)	0.5 (6.8)	0.0 (0.0)	0.7 (8.2)	0.4 (6.3)	0.1 (1.3)
Avg % Other	2.0 (12.3)	2.6 (12.3)	7.5 (19.8)	1.3 (10.1)	1.2 (8.9)	1.2 (9.7)	1.4 (11.6)	1.4 (11.0)	0.8 (8.3)	1.5 (9.5)
Avg % OtherEffect	14.6 (30.7)	14.9 (28.2)	11.5 (25.1)	9.0 (25.6)	16.2 (31.0)	13.8 (31.4)	9.2 (28.4)	14.0 (33.1)	20.3 (39.2)	13.7 (29.1)
Avg % PersonOrg	3.1 (14.9)	3.3 (13.3)	3.3 (14.3)	3.7 (15.4)	3.2 (14.2)	1.7 (11.9)	0.5 (7.0)	1.5 (11.6)	0.4 (6.3)	1.4 (8.7)
Avg % Prognosis	8.9 (24.8)	10.0 (22.5)	8.0 (21.2)	10.5 (27.1)	7.8 (21.7)	8.9 (25.5)	5.7 (23.2)	8.0 (25.8)	5.0 (21.5)	6.5 (19.4)
Avg % Susceptibility	10.2 (25.9)	15.3 (26.2)	14.4 (27.4)	16.0 (32.9)	17.6 (30.3)	7.0 (22.7)	4.7 (20.6)	6.1 (22.7)	7.1 (24.8)	16.7 (29.7)
Avg % Who	0.4 (6.0)	0.4 (5.0)	0.0 (0.0)	0.4 (4.2)	0.2 (3.1)	0.2 (3.6)	0.5 (7.0)	0.1 (2.6)	0.0 (0.0)	0.0 (0.0)
Avg % What	20.9 (36.8)	7.8 (21.3)	8.5 (21.7)	17.1 (32.5)	6.4 (20.7)	26.9 (40.4)	36.3 (47.5)	37.3 (45.8)	35.7 (46.6)	6.7 (20.4)
Avg % When	0.6 (6.8)	0.6 (5.9)	0.1 (1.1)	0.3 (5.1)	0.4 (5.5)	3.1 (15.7)	3.9 (19.3)	1.3 (10.9)	0.8 (9.0)	0.2 (4.2)
Avg % Where	1.4 (11.0)	0.4 (4.7)	0.0 (0.7)	1.7 (11.5)	0.6 (6.5)	0.1 (3.5)	0.0 (0.0)	0.4 (5.8)	0.8 (8.9)	0.2 (3.1)
Avg % Why	2.7 (14.9)	0.8 (7.1)	0.9 (6.7)	0.2 (3.4)	0.7 (7.2)	1.6 (12.0)	0.0 (0.0)	1.8 (12.9)	2.5 (14.0)	0.9 (8.0)
Avg % How	9.9 (27.4)	3.9 (15.1)	3.8 (14.4)	11.5 (27.3)	2.8 (14.1)	8.9 (26.1)	12.9 (33.3)	13.2 (32.4)	15.3 (35.6)	3.2 (16.1)
Avg % Non-WH	64.2 (43.1)	86.1 (27.3)	86.7 (26.2)	68.8 (40.9)	88.9 (26.6)	59.2 (44.5)	46.4 (49.2)	45.9 (47.0)	44.8 (48.2)	88.8 (26.9)

Numbers in parentheses are standard deviations. Percentages are out of 100.

consumers use fewer medical terms (MeSH: 7.0–15.9% versus 15.2–24.8%; SNOMED-CT: 16.6–23.2% versus 23.9–29.0%; CHV: 21.1–28.9% versus 30.4–41.8%). The ratio of SNOMED-CT to CHV terms is remarkably consistent for consumers (0.78–0.81:1) and is similar amongst professionals (0.77–0.81:1), except for JFPQ, which has a ratio of 0.61:1.

Finally, health questions can be compared based on their distributions of all UMLS semantic types to get a sense of the similarity of medical content. Appendix C shows the full distribution for each corpus; Table 3 summarizes the similarities using Jensen-Shannon divergence (0.0 indicates complete similarity; 1.0 indicates no similarity). The average divergence between the consumer corpora is 0.02489. The most dissimilar corpus is GARD, and if this corpus is removed the average divergence is just 0.00843. Professional corpora, however, are far less similar to each other, with an average divergence of 0.03812. The average professional divergence is actually greater than the divergence between all consumer questions and all professional questions, suggesting that professional corpora vary substantially in the types of medical concepts they contain.

Question Decomposition

Consumer questions tend to have around twice as many subquestions (1.7–2.4 versus 1.0–1.7). Across resources, the average number of consumer subquestions is remarkably consistent despite differences in question length. Instead, the longer questions contain more background information (1.2–4.5 sentences) and more ignore (nonpertinent) sentences (up to 0.5). Professional questions largely lack

Table 3. Jensen-Shannon divergence between the UMLS semantic type distributions of the consumer corpora, professional corpora, and combined consumer and professional corpora

		YANS			
Consumer corpora	WEBC	0.00949	WEBC		
	DSPR	0.00687	0.00881	DSPR	
	GARD	0.04729	0.04874	0.05485	GARD
	NLMC	0.00726	0.00614	0.01200	0.04741

		PHST			
Professional corpora	JFPQ	0.02519	JFPQ		
	CLIQ	0.00608	0.03308	CLIQ	
	PMOT	0.03558	0.03975	0.04576	PMOT
	NLMP	0.03287	0.06313	0.02861	0.07111

		CONS
Combined corpora	PROF	0.03363

background and ignore sentences, with the exception, again, of NLMP.

The question types indicate many differences between consumer and professional questions. Point-of-care questions (CLIQ and PMOT) are more interested in causes (7.0%, 9.1%). Consumer questions are less concerned about diagnosis (4.2–7.5% versus 7.4–11.0%), but more interested in general information (10.4–19.2% versus 4.4–15.7%). Both are very interested in management, markedly more for professionals (23.9–31.8% versus 33.1–57.5%). Consumers are more interested in manifestation (symptom) questions (1.3–3.3% versus 0.9–1.3%) and person/organization questions (3.1–3.7% versus 0.4–1.7%) that commonly ask for doctor or hospital information. Finally, ignoring NLMP, consumers are more interested in susceptibility information (10.2–17.6% versus 4.7–7.1%).

The wh-words give an indication of how questions are expressed, such as prototypical questions ("*What are the symptoms of . . .*," "*How do you treat . . .*") versus validation questions ("*Could nausea be a symptom. . .*," "*Is. . . a useful treatment?*"). For both consumers and professionals, use of questions that do not start with wh-words is quite high (64.2–88.9% versus 44.8–88.8%). When a wh-word is used, it is generally *what* or *how*. Within the consumer corpora, YANS and GARD have similar percentages of non-wh questions (64.2%, 68.8%), while the others are similarly close (86.1–88.9%).

Topics

Word clouds representing LDA topics built from UMLS terms are shown in Figures 2 and 3. The word clouds for the word-based LDA are provided in the supplemental data. The word clouds show sizable differences in content across consumer and professional questions as a whole. To compare content, a medical librarian (not an author) labeled each of the 100 topics with 4 categories that were commonly seen in the data: sexual health, cancer, medications, and diagnostic tests. A category was assigned if at least 2 of the top 25 topic words were related to the category. This analysis revealed that consumers (sexual health: 11 topics; cancer: 5; medications: 10; tests: 2) were far more likely to discuss sexual health and cancer, but much less likely to discuss medications or diagnostic tests relative to professionals (sexual health: 6; cancer: 1; medications: 31; tests: 6). Sexual health questions were disproportionally discussed in a single corpus (DSPR), while three-quarters of non-NLMP professional topics discussed medications.

Classification

Table 4 shows the results of the logistic regression model. Appendix D contains more details. According to the model, the language model features (open-domain versus medical) are the most discriminative, with the readability features being the second most discriminative and the semantic types being the least discriminative. The lack of weight given to the semantic types is curious, but the semantic types are likely subsumed by the medical concepts present, or absent, in the language models.

DISCUSSION

The above results show that consumer questions differ from professional questions in both form and content, which are affected by the particular resource. Previous work has shown that professional questions are shorter,[14] but we have demonstrated that professionals ask more succinct questions: fewer sentences, fewer subquestions, and less background information. Compare the WEBC question about bipolar disorder in Figure 1(a) to the PHST question in Figure 1(b). The professional question includes only the most relevant information in a qualifying clause ("*when combined with. . .*"). In contrast, consumer questions are filled with background information, even on QA websites with shorter questions. This stands in contrast to consumer search logs, where background information is difficult to include. Further, it is likely that one of the primary motivations for consumers posting questions online is that searches fail because online consumer resources are not sufficiently expressive: they are not designed to enumerate all possible symptoms and disease relationships commonly found in the background information of consumer questions. This would lead consumers to wonder whether the general-purpose resources are applicable to their case, or whether their details form an exception.

The traditional readability metrics indicate that consumer questions are easier to read. Looking at the 2 bipolar questions in Figure 1, however, might force one to conclude otherwise. This is a fundamental weakness with these readability metrics that base their scores only on word and sentence length. They do not account for case, orthographic, and grammatical errors, much less high-level coherence, all of which affect textual readability. Nevertheless, they reflect the reduced vocabulary and shorter sentences in the questions, which do improve readability. Regarding misspellings, other work has discussed the widespread presence of misspellings in consumer questions,[11] but our analysis allows us to determine that these are largely misspelled medical terms. Appendix E contains the 25 most frequent misspellings for each corpus, the vast majority of which are medical terms. This has important implications for systems that correct spelling in consumer text.[31,32]

Another distinction between consumer and professional language can be seen in the language model results. While we expected consumer questions to be closer to the open-domain language model, we did not expect that every consumer corpus would be favored by the open-domain model while every professional corpus would be favored by the medical model. Such a consistent result points to the oftentimes stark difference in language use between consumers and professionals. This even held true for GARD, which discusses diseases that are not commonly discussed in the open domain, thus indicating that professionals use a linguistic style that goes beyond disease terminology (eg, other medical terms, specialist jargon, clinical

Figure 2: Word clouds derived from 10-topic LDA using UMLS terms from the consumer corpora. (A high-definition image is available in the electronic version of this article.)

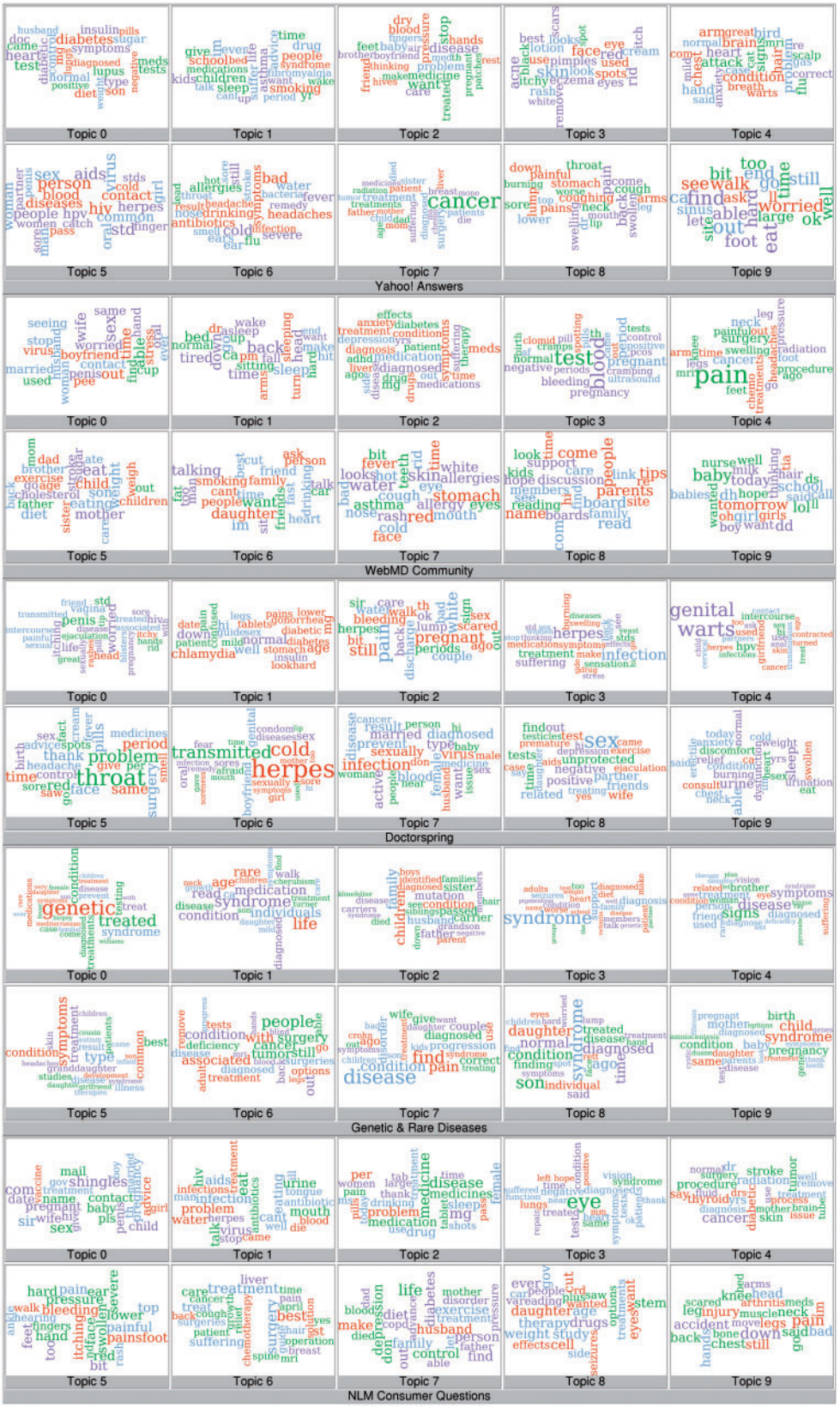

Figure 3: Word clouds derived from 10-topic LDA using UMLS terms from the professional corpora. (A high-definition image is available in the electronic version of this article.)

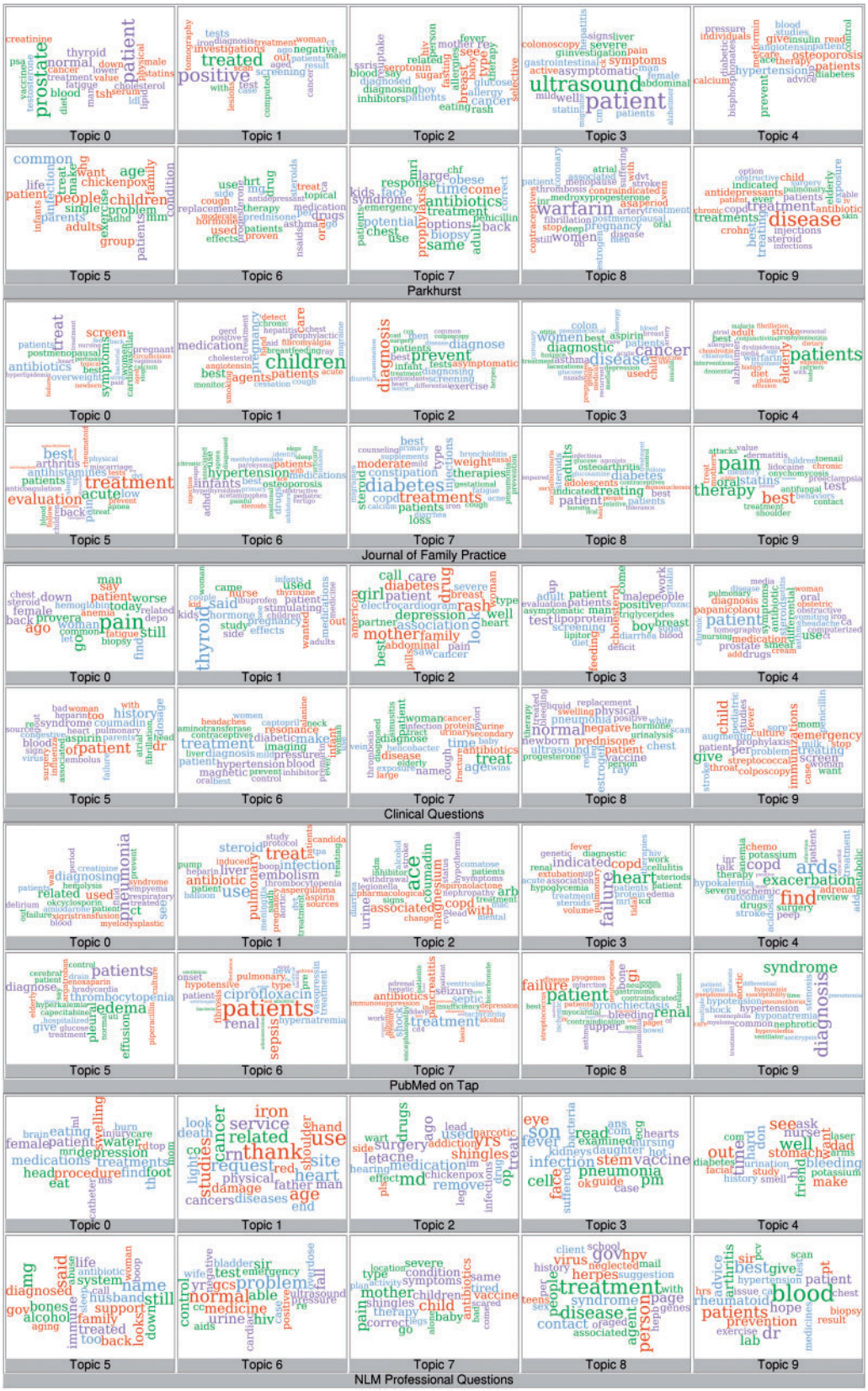

Table 4. Sum of logistic regression weights (absolute values) for all the features in the indicated feature sets

Weight Sum	Feature Type
0.05317	Lexical
0.05822	Readability
0.06647	Language Model
0.00634	Semantic Types
0.04317	Question Decomposition

Individual features are the same as the metrics from Tables 1 and 2. Individual weights are averaged over a 10-fold cross-validation. See supplemental Appendix D for individual feature results.

abbreviations, and syntactic structures). This has important implications for consumer NLP systems that utilize language models trained on professional text.[33]

While we are primarily concerned with consumers here and are utilizing professional corpora largely as a means of contrast, it is nonetheless interesting to note the differences among professional question corpora. As seen in Table 3, professional questions have greater semantic diversity than consumer questions. The distribution of semantic types (Appendix C) indicates that consumers discuss more anatomy and findings (characteristic of symptoms) and diseases receive approximately the same attention, while professionals discuss more laboratory and treatment procedures. Given the substantial differences between NLMP and the other professional corpora (and to a lesser extent NLMC and the other consumer corpora), some analysis is also merited on how the NLM corpora differ from the other corpora. First, unlike the others, the NLM questions are private, and thus their language tends to be more casual. The NLMC consumers are more willing to discuss personal and identifying details. The NLMP professionals are self-identified. While it is possible a large number of users misrepresented themselves, our analyses (eg, topics) suggest this is not the case. Instead, 2 major differences seem clear: (1) Unlike the other professionals, the NLMP professionals are more international and speak poorer English (Appendix F) shows the distribution of countries). In contrast, the other professionals are mostly from the United States and Canada. (2) The other professional corpora contain questions posed almost exclusively by physicians. The NLMP corpus includes many other types of health professionals, who are potentially more similar to consumers.

It is important to note the differences between consumer questions across resources, which has several possible causes. While previous work analyzing consumer text focused on a specific resource or subdomain,[9–12] our analysis spanned several different resources. We hope this makes our analysis more generalizable. Online resources are often viewed as communities, and thus form their own conventions that frequent users intentionally or unintentionally gravitate to. Some communities might encourage shorter or more detailed questions, while others might insist on proper spelling and grammar. Some communities might organically emphasize certain topics (eg, WebMD's substantial number of pregnancy-related questions), while others might be intentionally restrictive (eg, GARD's focus on genetic and rare diseases). We also suspect some communities might be better suited for younger audiences (eg, Yahoo! Answers[34]).

The intended audience should have a large impact on which resource to utilize. The NLM questions frequently contain private and identifying details not present in other question types. Another resource, Doctorspring.com, requires a fee, but comes with the benefit of the question being answered by a trained physician. It is thus interesting to note the preponderance of sexual health questions in this resource, suggesting a stratification effect where consumers view certain types of health problems (eg, sexually transmitted diseases) as worth the fee in return for more authoritative answers. It seems clear that consumers choose online resources based on a variety of causes: demographics, community, privacy, authority, and health topic.

Implications for automated question answering

The results presented in this study show that the difference between automated consumer and professional health QA systems should be more than a different terminology or answer corpus. Many QA systems are designed around the Ely questions (part of CLIQ), which are different from consumer questions in more than terminology and expected answers. A consumer QA system needs to handle longer questions, with more background information, and cannot assume a single specific question, as often multiple questions are asked at once. Consumers ask different types of questions than professionals, some types of which are hardly found at all in professional question corpora. One advantage is that consumer questions more closely resemble open-domain text, and thus open-domain NLP tools might prove more useful. In summary, a consumer health QA system should be designed with all these considerations from the start, instead of naively adapting a professional QA system.

Limitations

Despite dozens of measurements across 10 diverse corpora, this study still has 2 key limitations that could impact its conclusions. First, it is impossible to know how well the results from these corpora generalize to other online health questions. We were limited to those sites of which we were aware and whose data we could access. As is typical with informatics, completely new data sources would likely yield moderately different results at a minimum. Second, the automatic NLP methods certainly perform worse than a trained expert. We are unaware if the NLP methods are systematically biased in any way that may skew the results other than as described above.

CONCLUSION

Consumers extensively use various types of online resources to support their health decisions. Our study focuses on personal information-seeking beyond online searches and results in health-information needs explicitly stated as questions asked to peers and professionals. We demonstrate the various ways in which consumer and professional questions differ, which is important in order to guide resource construction as well as automatic consumer aids (eg, automatic QA). Our results show that consumers provide different amounts of background information and formulate questions differently depending on the particular resource. The choice to utilize a particular resource can be guided by various aspects of the consumer's case as well as the expected responder. Further, while there is great variation among consumer resources, it is likely not as great as the variation among professional resources. All of this reinforces a need for a variety of online health resources, as well as a need for informatics solutions to connect consumers with those resources beyond the use of standard search engines.

FUNDING

This work was supported by the National Institutes of Health (1K99LM012104), as well as the intramural research program at the US National Library of Medicine, National Institutes of Health.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

K.R. and D.D.F. designed the study. K.R. acquired the data and performed the implementation. K.R. and D.D.F. analyzed the results and wrote the paper.

REFERENCES

1. Lewis D, Eysenbach G, Kukafka R, *et al. Consumer Health Informatics: Informing Consumers and Improving Health Care*. Springer-Verlag New York; 2005.
2. Sadasivam RS, Kinney RL, Lemon SC, *et al.* Internet health information seeking is a team sport: Analysis of the Pew Internet Survey. *Int J Med Inform.* 2013;82(3):193–200.
3. Del Fiol G, Workman TE, Gorman PN. Clinical questions raised by clinicians at the point of care: a systematic review. *JAMA Intern Med.* 2014;174(5):710–718.
4. White RW, Horvitz E. From health search to health care: explorations of intention and utilization via query logs and user surveys. *J Am Med Inform Assoc.* 2014;21:49–55.
5. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc.* 2006;13(1):24–29.
6. Smith CA, Stavri PZ, Chapman WW. In Their Own Words? A Terminological Analysis of E-mail to a Cancer Information Service. *AMIA Annu Symp Proc.* 2002:697–701.
7. Zeng QT, Kogan S, Plovnick RM, *et al.* Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. *Int J Med Inform.* 2004;73:45–55.
8. Fox S. Peer-to-peer healthcare. *Pew Internet & American Life Project.* 2011.
9. White MD. Questioning Behavior on a Consumer Health Electronic List. *Library Quart.* 2000;70(3):302–334
10. Oh JS, He D, Jeng W, *et al.* Linguistic characteristics of eating disorder questions on Yahoo! Answers: content, style, and emotion. In *Proceedings of the American Society for Information Science and Technology*, 2013;50:1–10.
11. Zhang Y. Contextualizing Consumer Health Information Searching: an Analysis of Questions in a Social Q&A Community. *ACM International Health Informatics Symposium.* 2010:210–219.
12. Slaughter LA, Soergel D, Rindflesch TC. Semantic representation of consumer questions and physician answers. *Int J Med Inform.* 2006; 75:513–529.
13. Lou J, Zhang G-Q, Wentz S, *et al.* SimQ: real-time retrieval of similar consumer health questions. *J Med Internet Res.* 2015;17(2):e42.
14. Liu F, Antieau LD, Yu H. Toward automated consumer question answering: Automatically separating consumer questions from professional questions in the healthcare domain. *J Biomed Inform.* 2011;44(6):1032–1038.
15. Surdeanu M, Ciaramita M, Zaragoza H. Learning to rank answers on large online QA collections. *Assoc Comput Linguistics.* 2008:719–727.
16. Ely JW, Osheroff JA, Ebell MH, *et al.* Analysis of questions asked by family doctors regarding patient care. *BMJ.* 1999;319(7206):358–361.
17. Ely JW, Osheroff JA, Ferguson KJ, *et al.* Lifelong self-directed learning using a computer database of clinical questions. *J Fam Practice.* 1997;45(5):382–388.
18. D'Alessandro DM, Kreiter CD, Peterson MW. An Evaluation of Information-Seeking Behaviors of General Pediatricians. *Pediatrics.* 2004;113:64–69.
19. Hauser SE, Demner-Fushman D, Jacobs JL, *et al.* Using wireless handheld computers to seek information at the point of care: an evaluation by clinicians. *J Am Med Inform Assoc.* 2007;14(6):807–815.
20. Gunning R. *The Technique of Clear Writing.* McGraw-Hill; 1952.
21. Flesch R. A new readability yardstick. *J App Psychol.* 1948;32:221–233.
22. Kincaid JP, Fishburne RP, Rogers RL, *et al.* Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Research Branch Report 8-75. Chief of Naval Technical Training: Naval Air Station Memphis; 1975.
23. Stolcke A, Zheng J, Wang W, *et al.* SRILM at Sixteen: Update and Outlook. *IEEE Automatic Speech Recognition and Understanding Workshop.* 2011.
24. Parker R, Graff D, Kong J, *et al.* English Gigaword Fourth Edition. *The LDC Corpus Catalog*, 2009.
25. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med.* 1993;32(4):281–291.
26. Stearns MQ, Price C, Spackman KA, *et al.* SNOMED Clinical Terms: Overview of the Development Process and Project Status. *AMIA Annu Symp Proc.* 2001:662–666.
27. Roberts K, Kilicoglu H, Fiszman M, *et al.* Decomposing Consumer Health Questions. *BioNLP Workshop.* 2014:29–37.
28. Roberts K, Kilicoglu H, Fiszman M, *et al.* Automatically Classifying Question Types for Consumer Health Questions. *AMIA Annu Symp Proc.* 2014:1018–1027.
29. Roberts K, Masterton K, Fiszman M, *et al.* Annotating Question Types for Consumer Health Questions. *Building and Evaluating Resources for Health and Biomedical Text Processing.* 2014.
30. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res.* 2003;3:993–1022.
31. Crowell J, Zeng Q, Ngo L, *et al.* A Frequency-based Technique to Improve the Spelling Suggestion Rank in Medical Queries. *J Am Med Inform Assoc.* 2004;11(3):179–184.
32. Kilicoglu H, Fiszman M, Roberts K, *et al.* An Ensemble Method for Spelling Correction in Consumer Health Questions. *AMIA Annu Symp Proc.* 2015: 727–736.
33. Patrick J, Sabbagh M, Jain S, *et al.* Spelling correction in Clinical Notes with Emphasis on First Suggestion Accuracy. *Building and Evaluating Resources for Biomedical Text Processing.* 2010.
34. Kucuktunc O, Cambazoglu BB, Weber I, *et al.* A Large-Scale SentimentAnalysis for Yahoo! Answers. *ACM International Conference on Web Search and Data Mining.* 2012:633–642.

AUTHOR AFFILIATIONS

Lister Hill National Center for Biomedical Communications, US National Library of Medicine, National Institutes of Health 8600 Rockville Pike, Building 38A/1003H Bethesda, MD, 20894, USA