Analysis of Online Information Searching for Cardiovascular Diseases on a Consumer Health Information Portal

Ashutosh Jadhav, MS[*1], Amit Sheth, PhD[1], Jyotishman Pathak, PhD[2]
[1]Knoesis Center, Wright State University, Dayton, OH; [2]Mayo Clinic, Rochester, MN

ABSTRACT

Since the early 2000's, Internet usage for health information searching has increased significantly. Studying search queries can help us to understand users "information need" and how do they formulate search queries ("expression of information need"). Although cardiovascular diseases (CVD) affect a large percentage of the population, few studies have investigated how and what users search for CVD. We address this knowledge gap in the community by analyzing a large corpus of 10 million CVD related search queries from MayoClinic.com. Using UMLS MetaMap and UMLS semantic types/concepts, we developed a rule-based approach to categorize the queries into 14 health categories. We analyzed structural properties, types (keyword-based/Wh-questions/Yes-No questions) and linguistic structure of the queries. Our results show that the most searched health categories are 'Diseases/Conditions', 'Vital-Sings', 'Symptoms' and 'Living-with'. CVD queries are longer and are predominantly keyword-based. This study extends our knowledge about online health information searching and provides useful insights for Web search engines and health websites.

INTRODUCTION

Since the last decade, Internet literacy and the number of Internet users have increased exponentially. With the growing availability of online health resources, consumers are increasingly using the Internet to seek health related information[1,2]. Online health resources are easily accessible and provide information about most of the health topics. These resources can help non-experts to make more informed decisions and play a vital role in improving health literacy. According to the Center for Disease Control and Prevention (CDC) in the United States, CVD is one of the most common chronic diseases and the leading cause of death (1 in every 4 deaths) for both men and women[3]. CVD is common across all socioeconomic groups and demographics including age groups, genders, and ethnicities. Most of the CVDs require lifelong care and the patient is in charge of managing the disease through self-care (such as diet, exercise, and other health lifestyle choices)[4]. Prior studies[4,5] have shown that online resources are a "significant information supplement" for the patients with chronic conditions. As the percentage of people suffering from CVD is very high, the number of people using the Internet to search and learn about CVD is also large[4,5].

One of the most common ways to seek online health information is via Web search engines such as Google, Bing, Yahoo!, etc. According to the Pew Survey[1], approximately 8 in 10 online health inquiries start from a Web search engine. Therefore, studying CVD related search logs can help us to understand what health topics Online Health Information Seekers (OHIS) search for ("information need") and how do they formulate search queries ("expression of information need"). Such knowledge can be applied to improve the health search experience as well as to develop more advanced next-generation knowledge and content delivery systems. Although chronic diseases affect a large population, very few prior studies have investigated online health information searching exclusively for chronic diseases and especially for CVD. In this study, we address this knowledge gap in the community by performing a comprehensive analysis on a significantly large corpus of 10 million CVD related search queries. These queries are submitted from Web search engines and directed OHISs to the Mayo Clinic's consumer health information portal[6].

One of the contributions of this paper is the demonstration of the effectiveness of UMLS MetaMap[7] as well as UMLS semantic types and concepts for customized categorization. We implemented a rule based categorization approach (with Precision: 0.8842, Recall: 0.8607 and F1 Score: 0.8723) based on the UMLS semantic types and UMLS concepts. Using our approach, we categorized 92% of the 10 million CVD related search queries into 14 "consumer oriented" health categories. Additionally, we analyzed the structural properties of the queries (length of the search queries, usage of search query operators and special characters in the search queries), types of search queries (keyword-based, Wh-questions, Yes/No questions), and misspellings in the search queries. As the linguistic structure of the search queries has implications on information retrieval using Web search engines[8], we have also analyzed basic linguistic characteristics of the CVD search queries.

---

[*] This work was done during author's internship at Mayo Clinic, Rochester, MN, United States.

As per our analysis, the top searched health categories for CVD are 'Diseases and Conditions', 'Vital Sings', 'Symptoms' and 'Living with'. CVD search queries are longer and are predominantly keyword based. CVD queries have few search query operators, special characters, and spelling mistakes. This study provides useful and interesting insights for online health information seeking in chronic diseases and particularly in CVD. Such knowledge gives us better understanding about how OHIS search for CVD information and their information needs, which can be utilized to improve the health information search process.

Related work

Many previous studies have investigated online health information searching behavior, primarily using focus group studies, user surveys, and by analyzing health-related Web search query logs. In the studies[9-11] based on focus group and user surveys, researchers have analyzed online health searching characteristics such as how people use the Internet for health information searching, their demographic information (age, gender, education level, etc.), devices/Web search engines used for the searching, online health searching in the case of specific health conditions, and age-groups. Although these studies provided important insights, their main limitation was the inclusion of a limited number of participants (ranging from 100 to 2000 people), which may not cover population-level diversity and represent all socioeconomic groups. Researchers have analyzed web search logs from the health domain with diverse objectives, such as health/epidemic surveillance[12], PubMed usage[13], and online health information searching[14-19]. The studies focusing on online health information searching, have studied a variety of aspects of health query logs, such as health query length, changes in the health behavior with type of disease[17], effect of device used for health information searching[18] and changes in online health search patterns with disease escalation from symptoms to serious illness[19].

Previous studies in chronic diseases have looked at different facets of the diseases such as how education, income, and occupation contribute to risk factors for cardiovascular diseases[20]; use of information technology to improve the management of chronic diseases which includes home monitoring of vital signs for patients with chronic diseases[21]. Ayer et al.[4] studied the relationship between chronic illness, use of the Internet for health information, and change in health behavior. The study suggests that the use of the Internet empowers patients' in the management of their chronic conditions resulting in an increased ability to make informed decisions about their health. Lorig et al.[22] indicated the effectiveness of Internet based chronic disease self-management. Although chronic diseases affect a large population and prior studies have highlighted the usefulness of the online health resources for patients with chronic diseases, very few studies[4] have investigated how and what OHIS search for chronic diseases and especially for CVD. A focused study on chronic diseases, such as the CVD use-case presented in this paper, helps us to study details about online health searching for chronic diseases that may not be revealed through analyzing online health searching in general. Such knowledge can be applied to improve online health information systems, to promote health literacy and to accomplish a more balanced approach to wellness and prevention of the chronic diseases.

MATERIALS AND METHODS

Data Source: In this study, we have collected CVD-related search queries originating from Web search engines (such as Google and Bing) that direct OHISs to Mayo Clinic's consumer health information portal[6] (MayoClinic.com), which is one of the top online health information portals within the United States. The MayoClinic.com portal provides up-to-date, high-quality online health information produced by professional writers and editors. Our recent Web analytics statistics indicate that the MayoClinic.com portal is on average visited by millions of unique visitors every day and around 90% of the incoming traffic is originated from Web search engines. This significant traffic to the portal provides us with an excellent platform to conduct our study.

Dataset Creation: The MayoClinic.com Web Analytics tool (IBM Netinsight on Demand) keeps detailed information about Web traffic such as input search query, time of visit, and landing page. MayoClinic.com has several CVD-related webpages that are organized by health topics and disease types. Using the Web Analytics tool, we obtained 10 million CVD-related anonymized search queries originating from Web search engines that "land on" CVD webpages within MayoClinic.com and are related to CVD. These queries are in English language and are collected between September 2011-August 2013. Our final analysis dataset consists of 10,408,921 CVD related search queries, which is a significantly large dataset for a single class of diseases.

Data Analysis: We performed the following analysis on the CVD related search queries: 1) Top search queries associated with CVD; 2) Semantic analysis: categorization of the queries into health categories using UMLS MetaMap; 3) Structural analysis: length of the search queries in number of words and number of characters, usage of search query operators (such as 'and' and 'or') and special characters in the search queries; and 4) Textual analysis:

types of search queries (keyword based, Wh-questions, Yes/No questions), misspellings in the queries, and linguistic structure of the search queries (part-of-speech analysis).

1. Top CVD search queries: We selected the 20 most frequently searched queries from the analysis dataset.

2. Categorization of the queries into health categories using UMLS MetaMap

2.1. Selection of health categories: There are many possible health categories of interest. In this work, we selected 14 health categories **(Table. 1)** that are "consumer oriented" as well as can reveal details about what OHISs generally search for in the context of CVD. Here, we define "consumer oriented" health categories as categories that are easily understandable for a non-expert, lay population. While selecting the health categories, we studied the health categories on popular health websites (e.g., Mayo Clinic, WebMD, etc.) and the types of information frequently mentioned along with CVD search queries, e.g. vital signs (blood pressure, heart rate), age groups (infants, adult, elder), etc. Note that there can be possible overlaps between some of the health categories, for example 'Drugs and Medications' can be considered as a part of 'Treatment', but in our analysis we considered both as a separate health categories in order to study search traffic for each category separately. These categories and the categorization scheme **(Table. 1)** is reviewed and verified by the Mayo Clinic clinicians and domain experts.

Table 1. List of health categories and their description with examples

Categories	Description and Examples
Symptoms	Search queries related signs and symptoms, e.g. Stroke symptoms, heart palpitations with headache, home remedies for heart murmur, heartburn vs heart attack symptoms
Causes	Search queries related to cause/reasons for various CVD conditions, symptoms, e.g. causes of an elevated heart rate, heart failure reasons, morning hypertension causes
Risks and Complication	Search queries related to risk and complications, e.g. risks of pacemaker, risk factors to hypertension, complications of bypass surgery, heart ablation surgery risks
Drugs and Medications	Search queries related to drugs and medications, e.g. dextromethorphan blood pressure, medications pulmonary hypertension, tylenol raise blood pressure, ibuprofen heart rate
Treatments	Search queries related to treatments, e.g. exercise for reducing hypertension, cardiac arrest treatments, dilated cardiomyopathy treatment, bypass surgery, cardiac rehabilitation
Tests and Diagnosis	Search queries related to tests and diagnosis, e.g. heart echocardiogram, diagnosis of vascular disease, ct scan for heart, test for cardiomyopathy, urinalysis in hypertension
Food and Diet	Search queries related to food and diet, e.g. what is cardiac diet, what foods lower blood pressure and cholesterol, red wine heart disease, alcohol and hypertension
Living with	Search queries related to control, management, cure and living with CVD, e.g. exercises to lower high blood pressure, controlling blood pressure naturally, cure for postural hypotension, lifestyle changes to lower hypertension, living with pacemaker, how to control cholesterol
Prevention	Search queries related to prevention, e.g. ways to prevent heart attack, preventing stroke, foods to avoid heart diseases, aspirin for prevention of stroke, foods that lower risk of heart disease
Side effects	Search queries related to side effects, e.g. blood pressure pills side effects, side effects of beta blockers for hypertension, coq10 bp side effects, side effects of pulmonary hypertension
Medical devices	Search queries related to medical device references, e.g. living with a pacemaker, using blood pressure cuff, pump for pulmonary hypertension, blood pressure monitor
Diseases and conditions	Search queries related to diseases and conditions, e.g. medications pulmonary hypertension, born with holes in heart, stroke tia symptoms, hypotension, heart attack in pregnancy
Age-group References	Search queries with references to age groups, e.g. cardiac defects in children, average heart rate for an adult, heart murmur in adult, hypertension in adolescents, heart murmurs in infants
Vital signs	Search queries with references to blood pressure, heart rate, pulse rate, temperature, heart beat (without high/low blood pressure as we considered them under 'Diseases and Conditions'), e.g. blood pressure 125/90, normal resting heart rate, can tylenol raise blood pressure, healthy heart rate chart, sleep and blood pressure

2.2. Mapping CVD search queries to UMLS semantic types and concepts: We performed semantic analysis on the CVD search queries by mapping all the search queries from the dataset to UMLS concepts and semantic types using UMLS MetaMap[7]. MetaMap is a tool for recognizing UMLS concepts in the text. For a given search query, MetaMap identifies one or more UMLS concepts, their semantic types, Concept Unique Identifiers (CUIs), and other details. UMLS incorporates variety of medical vocabularies and concepts, and maps each concept to semantic types. Thus using UMLS, we can understand 'semantics' or 'meaning' of the words and phrases. For example: in

search query 'heart attack red wine' semantics is used to understand that the query is about two separate phrases 'heart attack' and 'red wine' and not about 4 separate words 'heart', 'attack', 'red' and 'wine'. Moreover using semantics facilitated by UMLS, we can understand "heart attack" is a disease or syndrome (DSYN) and "red wine" is a "Food item" (FOOD). Refer to Table 2 for semantic type abbreviations used in this paper.

Table 2. List of health categories and their respective UMLS semantic types/concepts used categorization. **Abbreviations**: SOSY-Signs and Symptoms, ORCH-Organic Chemical, PHSU- Pharmacologic Substance, CLND-Clinical Drug, TOPP- Therapeutic or Preventive Procedure, FTCN-Functional Concept, CNCE-Conceptual Entity, DIAP- Diagnostic Procedure, LBPR-Laboratory Procedure, LBTR- Laboratory Test Result, FOOD-Food, MEDD-Medical Device, DSYN-Disease or Syndrome, AGGP-Age Group

Health Categories	UMLS Semantic Types (**ST**), UMLS Concepts (**CC**) and Keywords (**KW**)
Symptoms	**ST**: SOSY **CC:** symptoms, signs, heart murmur
Causes	**CC**: cause, reason
Risks & Complications	**CC**: risk, complications
Drugs and Medications	**ST**: ORCH|PHSU, CLND, PHSU **CC**: medication, medicine, drugs, dose, dosage, tablet, pill **KW**: meds (without CC: alcohol, caffeine, fruit, prevent)
Treatments	**ST**: TOPP, FTCN (treatment, surgery), CNCE (treatment), **CC**: remedy, remediate (without CC: prevention and 'Drugs and Medication' queries)
Tests and Diagnosis	**ST**: DIAP, LBPR, LBTR **CC**: Test, diagnosis (without ST: DIAP| TOPP, CC: alcohol, blood caffeine)
Food and Diet	**ST**: FOOD **CC:** caffeine, recipe, meal, menu, diet, eat, breakfast, lunch, dinner, alcohol, drink
Living with	**CC:** control, manage, reduce, lower, coping, cure, recover **KW:** living with, bring down, low down
Prevention	**CC:** prevent, avoidance, low risk
Side effects	**CC:** side effect **KW:** side effect
Medical devices	**ST**: MEDD
Diseases and conditions	**ST**: DSYN **CC:** arrhythmia, avascular necrosis, enlarged heart, hypotension, blood pressure low **KW**: heart damage
Age-group References	**ST**: AGGP
Vital signs	**CC:** blood pressure, heart rate, pulse rate, temperature, Heart beat (without high/low blood pressure as we considered them under 'Diseases and Conditions')

2.3. Categorization Approach (**Table 2**): We categorized the search queries into 14 health categories as following:

1) UMLS has 140 semantic types and some of them are directly mapped to health categories that we selected; for example, 'AGGP' (Age-group) semantic type is directly mapped to the 'Age group' category. In this case, we categorized all the search queries with semantic type 'AGGP' into the 'Age group' category.
2) For a few health categories (e.g., 'Test and Diagnosis') we utilized multiple semantic types ('DIAP', 'LBPR', 'LBTR'). In this case, we categorized all the queries with at least one semantic type ('DIAP', 'LBPR', 'LBTR') into the 'Test & Diagnosis' category.
3) For a few health categories (e.g., 'Food and Diet'), there are certain concepts that are closely associated with the health category are not mapped to the selected semantic type. In such cases, we utilized both semantic types and well as semantic concepts for the categorization. For example, 'FOOD' semantic type does not include concepts such as 'meal', 'menu', 'diet', 'recipe' and 'lunch' as they are not actually food items. We categorized all the search queries that have 'FOOD' semantic type or at least one concept from (meal', 'menu', 'diet', etc.) into the 'Food and Diet' category.
4) For a few health categories (e.g., 'Cause') there is no directly associated semantic type. In such cases, we utilized semantic concepts for the categorization. For example, we categorized all the search queries which have either 'Cause' or 'Reason' semantic concepts into the 'Cause' category.
5) For a few health categories (e.g., 'Living with'), apart from semantic concepts, we also considered the presence of keywords ('Living with') within the search query as 'Living with' is not a concept in the UMLS.
6) Few semantic types include some undesired concepts (in the context of our customized categorization, not in the terms of UMLS concept hierarchy). For example, semantic types 'ORCH\| PHSU' and 'PHSU' are associated with the 'Drugs and Medication' category. These semantic types include some concepts that are not considered as drugs to a consumer/lay population: caffeine, fruit, prevent, etc. In such cases, we do not categorize the search

queries with semantic types 'ORCH/PHSU' or 'PHSU' and with semantic concepts caffeine, fruit, prevent, etc. into the 'Drugs and Medication' category.

7) We also considered lexical variants, as well as partial matches, of some concepts for example: diagnose, diagnosis, test, testing, etc. A search query can be categorized into zero, one or more than one health category depending on the mapping of the query to UMLS concepts and semantic types. We empirically evaluated queries in each category and performed several iterations to evaluate the semantic type/concepts for each category then defined the categorization scheme **(Table 2)**.

2.4. Categorization Evaluation: We evaluated the performance of the categorization approach as following:

1) Gold standard dataset creation: We randomly selected 2000 search queries from the analysis dataset. Two domain experts manually annotated 2000 search queries by labeling one search query with zero, one, or more than one health category. The annotators first discussed and agreed upon the annotation scheme. To reduce the probability of human errors and subjectivity, the two annotators discussed together and annotated each query and created a gold standard dataset with 2000 search queries, which is further divided into training and testing datasets with 1000 search queries each. Training dataset is used to develop rule-based categorization approach.
2) Precision-Recall calculation: We categorized 1000 search queries from the testing dataset using the categorization approach as discussed in Section 2.3 and evaluated the categorization approach with respect to the gold standard dataset. Since we categorized search queries into 14 health categories, we also calculated Micro average Precision and Recall. Based on the evaluation, our categorization approach has very good Precision: 0.8842, Recall: 0.8642 and F-Score: 0.8723.
3) We also performed Precision and Recall analysis for each health category independently (Table 3) to check the performance of the categorization approach for individual health categories. The categorization approach works well for most of the categories while for a few categories the approach shows above average Precision/Recall. One observed reason that affected the Precision/Recall is multiple interpretations of the concepts that sometime may not be contextually correct, e.g. for the search query 'nuts good for your heart', MetaMap annotated 'nuts' as 'FOOD' as well as 'MEDD' (Nut - Medical Device Component or Accessory).

Table 3. List of health categories and their respective Precision and Recall

No	Categories	Precision	Recall	F1 Score	No.	Categories	Precision	Recall	F1 Score
1	Symptoms	0.9274	0.8042	0.8614	8	Living with	0.8659	0.9342	0.8988
2	Causes	0.8861	0.9859	0.9333	9	Prevention	0.8333	1.0000	0.9091
3	Risks and Complication	1.0000	1.0000	1.0000	10	Side effects	1.0000	1.0000	1.0000
4	Drugs and Medications	0.8582	0.9350	0.8950	11	Medical devices	0.8077	0.7500	0.7778
5	Treatments	0.7083	0.9444	0.8095	12	Diseases	0.9291	0.7751	0.8451
6	Tests and Diagnosis	0.6389	1.0	0.7797	13	Age-group References	1.0000	0.8889	0.9412
7	Food & Diet	0.9391	0.9558	0.9474	14	Vital signs	0.8872	0.8669	0.8769
Overall Micro Average **Precision (0.8842), Recall (0.8607)** and **F1 Score (0.8723)**									

3. Health query length: We calculated search query length by computing the number of words (separated by white space) and the number of characters (excluding white space) in the search queries.

4. Usage of query operators and special characters: In search queries, query operators ('and', 'or', 'not', etc.) are used to formulate complex queries. In this study, we have considered following operators: AND, OR, +, &, other (NOT, AND NOT, OR NOT, & NOT). Special characters are the characters apart from letters (a-z) and digits (0-9). The significance of special characters in health search query depends upon the usage of special characters in the medical domain. For example, OHISs may mention values in different formats, e.g., 2.3 ml, 40%, 17-19, 125/90 (for blood pressure) or $200 (for the cost of a drug or procedure). We analyzed the usage of search query operators and special characters in the CVD search queries based on their usage frequency in the search queries.

5. Misspellings in health queries: OHISs occasionally make spelling mistakes while searching for health information. To analyze the frequency of such errors, we used a dictionary-based approach. We first generated a dictionary of words using the Zyzzyva wordlist[23], the Hunspell dictionary[24], and its medical version (OpenMedSpell[25]), comprising a total of 275,270 unique words. We used this dictionary to check misspellings in the CVD search queries.

6. Type of Search queries: OHISs express their health information need by formulating search queries on Web search engines. OHISs can express their information need either by formulating search queries using keywords or asking questions (Wh-questions and Yes/No questions). For this analysis, we have considered the following Wh-questions (lexicon): 'What', 'How', '?', 'When', 'Why' and others ('Who' 'Where', 'Which'). Note that, although '?' does not come under Wh-questions category, we have included it for the simplicity. Yes/No questions are usually used to check some factual information, for example, whether coffee is bad for the heart. In this analysis, we have considered Yes/No questions that start with 'Can', 'Is', 'Does', 'Do', 'Are', and others ('Could' 'Should', 'Will', 'Would'). Using the lexicon for Wh-questions and Yes/No questions, we performed text analysis on the search queries to count the number of queries with Wh-questions and Yes/No Questions. Search queries that do not contain any questions (Wh- or Yes/No) are classified as Keyword-based. Additionally, for different Wh-questions and Yes/No questions, we computed their usage frequency in the search queries.

7. Linguistic analysis of health queries: Linguistic structure of the search queries has implications on information retrieval using Web search engines[8]. Thus we analyzed basic linguistic characteristics of the CVD search queries. We performed part-of-speech analysis on the search queries using Stanford's POS tagger[26]. For this analysis, we considered nouns, verbs, adjectives and adverbs. We mapped all the subtypes in part-of-speech (e.g. proper nouns, common nouns, compound nouns) to the main part-of-speech types (e.g. nouns). We analyzed usage of different part-of-speech types in the CVD queries based on their usage frequency in the search queries.

RESULTS

1. Top health queries: Most of the top search queries are related to major CVD diseases and conditions. At the same time, questions about blood pressure (high/low) and heart rate were also searched frequently (Table 4).

Table 4. Top search queries related to CVD

Top 1-5 Queries	Top 6-10 Queries	Top 11-15 Queries	Top 16-20 Queries
heart attack symptom	congestive heart failure	cardiomyopathy	echocardiogram
blood pressure chart	low blood pressure	heart palpitations	heart disease
how to lower blood pressure	stroke symptoms	blood pressure medication	orthostatic hypotension
heart rate	normal blood pressure	symptoms of stroke	heart healthy recipes
broken heart syndrome	high blood pressure symptoms	heat stroke	heart arrhythmia

2. Health categories: Based on **Table 5**, the most popular health categories while searching for CVD information are 'Diseases and Conditions' and 'Vital signs'. One in every two searches is related to either 'Diseases and Conditions' or 'Vital signs'. Due to close association of vital sings (such as blood pressure and heart rate) with CVD, OHIS might be searching it frequently. Other popular health categories that users search for includes 'Symptoms', 'Living with', 'Treatments', 'Food and Diet' and 'Causes'. Mostly due to the chronic nature of the CVD and as the patients are in charge of managing the disease with day-to-day care, many CVD patients might be searching for 'Living with' related search queries. As diet has a significant impact on the CVD, we observed large search traffic for 'Food and Diet' category. Many OHISs are also interested in learning about CVD 'Treatments', 'Medical Devices' (e.g. pacemaker), 'Drugs and Medication', and 'Cause'. Although CVD can be prevented with some lifestyle and diet changes, interestingly very few OHISs search for CVD 'Prevention'.

Table 5. Categorization of CVD search queries into 14 health categories

No	Health categories	Total Queries	Percentage Distribution	No	Health categories	Total Queries	Percentage Distribution
1	Diseases	4,232,398	28.66	8	Drugs and Medications	603,905	4.09
2	Vital signs	3,455,809	23.40	9	Causes	599,895	4.06
3	Symptoms	1,422,826	9.64	10	Tests & Diagnosis	344,747	2.33
4	Living with	1,178,756	7.98	11	Risks and Complication	277,294	1.88
5	Treatments	955,701	6.47	12	Prevention	136,428	0.92
6	Food and Diet	779,949	5.28	13	Age-group References	87,929	0.60
7	Med Devices	665,484	4.51	14	Side effects	25,655	0.17
					Total	14,766,776	100

Using our categorization approach, we categorized 92% of the 10 million CVD related queries into at least one health category (**Table 6**). Most of the queries (around 88%) are categorized into either one or two categories (Table 6). Very few CVD queries (4.28%) are categorized into 3 or more categories. Our approach did not categorize 8.13% of the queries into any health categories. After studying the uncategorized search queries, we found that there are few queries that do not fit into any of the selected 14 categories such as cardiac surgeon, cardiology mayo, video on cardiovascular, pediatric cardiology, and orthostatic.

Table 6. A search query can be categorized into zero, one, or more health categories. The table shows the distribution of search queries by number of health categories in which they are categorized.

Number of health Categories	Number of search queries	Percentage Distribution
0	845,744	8.13%
1	4,967,337	47.72%
2	4,149,803	39.87%
3	420,622	4.04%
4 and 5	25,415	0.24%
Total	10,408,921	100.00%

3. Health query length: Average search query length (**Figure 2**) for CVD is 3.88 words and 22.22 characters. Around 80% of the CVD search queries have 3 or more words. The analysis implies that, CVD search queries are longer than previously reported non-medical, as well as medical queries, as the average length for both of them is around 2.35 words[15,27]. This potentially indicates that OHISs describe their CVD information needs in more detail by adding relevant health context to the search query. Longer search queries also denote OHIS' interest in more specific information about the disease; subsequently OHISs use more words to narrow down to a particular health topic. Another possible reason for longer CVD search queries might be that even simple CVD related search queries have multiple words (Table 1, Table 4).

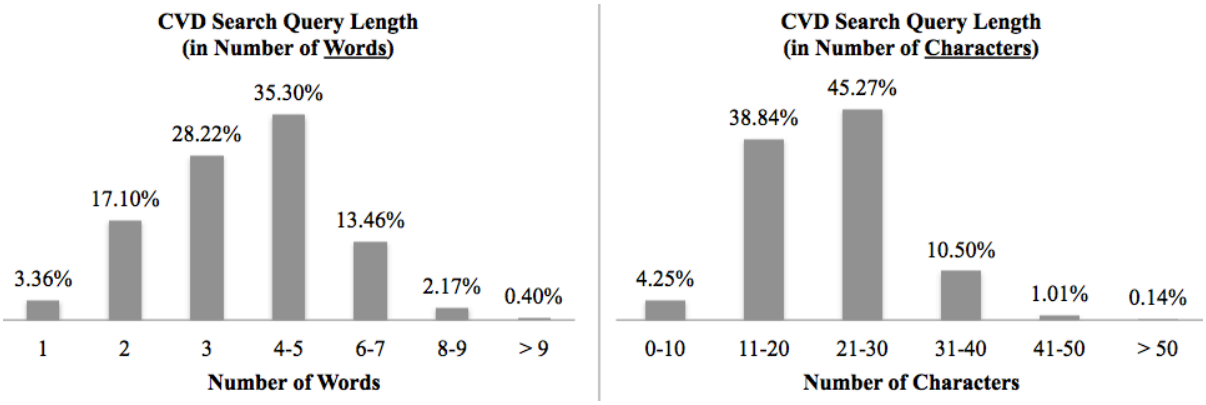

Figure 2. Distribution of length of search queries by number of words and number of characters

Table 7. Search query operators usage, special characters usage, and misspellings in the CVD search queries

Structural Analysis	Usage Frequency	Total Queries	Percentage Distribution
Number of Query Operators	0	10,011,257	96.18%
	>0	397,664	3.82%
Query Operators Usage	AND	366,117	92.07%
	+	7,115	1.79%
	OR	16,063	4.04%
	&	4,159	1.05%
	Other	4,210	1.06%
Special Characters	0	10,288,916	98.85%
	>0	120,005	1.15%
Spelling Mistake	0	10,075,665	96.80%
	>0	333,256	3.20%

4. Query operators usage, special characters usage, and misspellings in the CVD search queries (Table 7): Around 4% of CVD search queries use at least one query operator. 'AND' is the most popular operator (92%), followed by 'OR' (4%) and '+' (1.7%). Overall variations of 'and' (AND, &, +) operators comprise around 95% of operator usage in the search queries. OHISs formulate very few (1%) CVD search queries with special characters. In CVD search queries, 3.2% of the queries have at least one spelling mistake. Web search engine's "auto-completion" as well as "spelling correction" functionalities might be one reason to lower misspellings in the search queries.

5. Type of Search queries: As indicated by the analysis in **Figure 3**, OHISs predominantly formulate search queries using keywords (80%), though queries with Wh-Questions are also significant. Few queries (2.5%) are formulated as Yes/No type questions. In Wh-questions, OHISs mostly use "How" and "What" in the search queries and both of them generally signify that more descriptive information is needed. In Yes/No Questions, OHISs more often start the search queries with "does" "can" and "is".

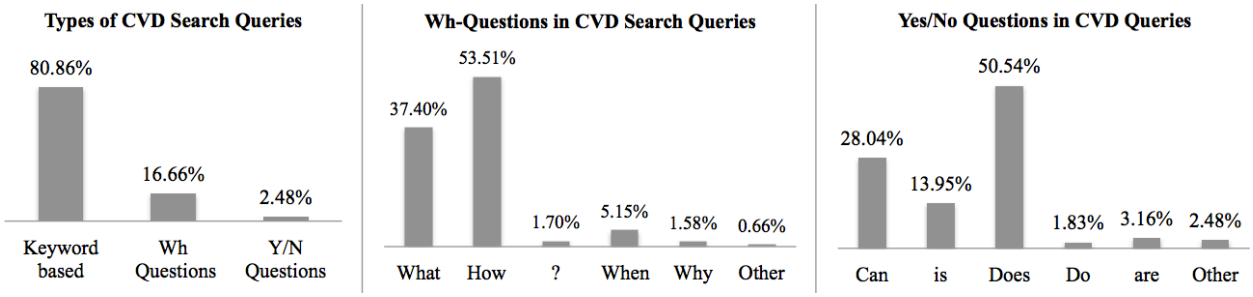

Figure 3. Types of search queries (how health information need is expressed) and distribution of Wh-Questions and Yes/No Questions based on frequency of their usage in the CVD search queries.

6. Linguistic analysis (Table 8): In health search queries, nouns typically denote entities like disease names, health categories, etc. Almost all the CVD search queries have at least one noun and most of the search queries (81.5%) have 2 or 3 nouns. A verb conveys an action or an occurrence, for example "how to control (verb) hypertension (noun)". Approximately 26% of the search queries have at least one verb. Adverbs are the words that modify a verb, an adjective and another adverb, while an adjective is a 'describing' word, giving more information about the object signified, for example "extremely (adverb) bad (adjective) heart (noun) pain (noun)." Very few search queries have at least one adverb and 47% of the queries have at least one adjective.

Table 8. Linguistic analysis (part-of-speech) on CVD search queries

Part-of-Speech	Usage Frequency	Total Queries	Percentage Distribution
Nouns	**0**	6,279	0.06%
	1	1,075,467	10.33%
	2	4,964,074	47.69%
	3	3,523,204	33.85%
	>3	839,897	8.07%
Verb	**0**	7,722,752	74.19%
	>0	2,686,169	25.81%
Adverb	**0**	10,151,641	97.53%
	>0	257,280	2.47%
Adjective	**0**	5,519,489	53.03%
	>0	4,889,432	46.97%

DISCUSSION

In this study, we analyzed a significantly large dataset of 10 million CVD related search queries in order to understand online health information searching for CVD. We implemented a rule based categorization approach (with Precision: 0.8842, Recall: 0.8607and F1 Score: 0.8723) using UMLS concepts/semantic types and categorized 92% of the 10 million CVD related search queries into 14 "consumer oriented" health categories. As per our analysis, the top searched health categories ("information needs") for CVD are 'Diseases and Conditions', 'Vital Sings', 'Symptoms', and 'Living with'. Other frequently searched CVD health categories are 'Treatments', 'Food and Diet', and 'Causes'. Most of the queries (around 88%) are categorized into either one or two health categories.

Even though CVD can be prevented with some lifestyle and diet changes, very few OHIS search for preventive health information. We found that use of MetaMap and UMLS concepts/semantic type to be a very good approach for categorization of the health related search queries as UMLS incorporates variety of medical vocabularies and concepts, and mapping of each concept to semantic types. However for customized categorization, we have to carefully select/eliminate UMLS semantic types and concepts considering the alignment of their scope with desired categories.

Our study reveals some interesting insights about structural and syntactic properties of CVD search queries. The average length of CVD search queries is longer than that of previously reported general search queries as well as medical search queries[15,27]. Our study implies OHISs may be interested in more specific information. Only 3.2% of the search queries contain at least one spelling mistake. The search engine's "auto-completion" feature, and spelling correction/suggestion functionality might be contributing to reduce misspelled words in search queries. Very few OHISs use search query operators and variations of 'and' (AND, &, +) operators comprise around 95% of operator usage in the search queries. OHISs formulate search queries primarily using keywords (around 80%), followed by Wh-Questions, and Yes/No Questions. In Wh-questions, OHISs mostly use 'What' and 'How' in the search queries, and both of them generally signify a need for more descriptive information, while search queries in the form of Yes/No questions indicate interest in factual information. Almost all the search queries have one noun. OHISs also use adjectives and verbs frequently in the search queries to add context to the topic of interest.

Following are some of the limitations of this study. The results of this study are derived from the analysis limited to CVD search queries from Web search engines that led users to MayoClinic.com. Even though Mayo Clinic web pages often ranked high in Web search engines, not all health information seekers visit MayoClinic.com. The focus of this study is limited to analysis of the search query log and we have not analyzed associated socioeconomic factors due to the anonymized nature of the data. To the best of our knowledge, there is not much research on understanding online health information searching for chronic diseases and especially for CVD. This study addresses this knowledge gap and extends our knowledge about online health information search behavior. The study provides interesting and valuable insights that can further be leveraged in multiple ways, such as:

1. Web search engines: to understand details about structural and linguistic characteristics of health search queries, search query complexity, and popular health categories in order to improve health information retrieval systems;
2. Websites that provide health information: to better understand an OHIS's health information need, and do a better organization of health information content;
3. Healthcare providers: to better understand their patients and their health information interest;
4. Healthcare-centric application developers: to better understand what and how OHISs search for and to build applications around consumer health information needs and priorities;
5. OHISs: we anticipate that this work will help empower OHISs in their quest for health information, and facilitate their health information search efforts by enabling the development of smarter and more sophisticated consumer health information delivery mechanisms.

In the future, we plan to leverage insights from this work to facilitate a better health search experience by developing more advanced next-generation knowledge and content delivery systems. Also, we plan to perform comparative analysis of major diseases to learn similarities and differences between them in online health information searching.

Conclusion

We presented a comprehensive analysis on CVD related search queries in order to understand what users search for ("information need") and how they formulate search queries ("expression of information need"). We found that using MetaMap and UMLS concepts/semantic type is a very good approach for categorization of health related search queries into health categories. The categorization approach can be reused for different set of health categories by defining new association rules**.** Distribution of the CVD queries by health categories indicates that, OHISs have most information needs for CVD related 'Diseases and Conditions', 'Vital Signs' 'Symptoms' and post CVD information ('Living with', 'Diet', 'Treatments, Drugs'). OHISs predominantly formulate search queries using keywords followed by Wh-Questions and Yes/No Questions. Almost all CVD search queries have at least one noun. A greater understanding of OHIS's needs may help us to accomplish the changes that will lead to improvement in online health information searching and a more balanced approach for health information intervention. This study extends our knowledge about online health information searching, and provides useful insights for Web search engines, health-centric websites and application developers. Finally, we anticipate that this work will help empower

OHISs in their quest for health information, and facilitate their health information search efforts by enabling the development of more advanced next-generation knowledge and content delivery systems.

Acknowledgement: We thank the Mayo Clinic Web Analytics team for their valuable contribution in data provision. This work is supported by the Mayo Clinic NIH Relief Fund Award (FP00068008). We acknowledge Rashmi Dusane for her help in this work.

References

1. Fox S, Duggan M. Health online 2013. Pew internet & American Life Project 2013.
2. Higgins O SJ, Barry MM, Domegan C. A literature review on health information seeking behaviour on the web: a health consumer and health professional perspective. In: ECDC, ed. Stockholm2011.
3. National Center for Health Statistics. Health, United States, 2010: With special feature on death and dying Hyattsville, MD. 2011.
4. Ayers SL, Kronenfeld JJ. Chronic illness and health-seeking information on the Internet. Health:. 2007;11(3):327-347.
5. Fox S, Duggan M. Who Lives with Chronic Conditions Pew internet & American Life Project. 2013.
6. Mayo Clinic's consumer health information website. http://www.mayoclinic.com/ Accessed March 9, 2014
7. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Paper presented at: Proceedings of the AMIA Symposium2001.
8. Croft WB et.al. Search engines: Information retrieval in practice. Addison-Wesley Reading; 2010.
9. Drentea P, Goldner M, Cotten S, Hale T. The association among gender, computer use and online health searching, and mental health. Information, Communication & Society. 2008;11(4):509-525.
10. Weaver III JB, Mays D, Weaver SS, Hopkins GL, Eroglu D, Bernhardt JM. Health information-seeking behaviors, health indicators, and health risks. American journal of public health. 2010;100(8):1520-1525.
11. Atkinson NL, Saperstein SL, Pleis J. Using the internet for health-related activities: findings from a national probability sample. Journal of Medical Internet Research. 2009;11(1).
12. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. 2009;457\(7232\):1012\-1014\.
13. Herskovic JR, Tanaka LY, Hersh W, Bernstam EV. A Day in the Life of PubMed: Analysis of a Typical Day's Query Log. Journal of the American Medical Informatics Association. 3// 2007;14(2):212-220.
14. Cartright M-A, White RW, Horvitz E. Intentions and attention in exploratory health search. SIGIR2011.
15. Spink A, Yang Y, Jansen J, et al. A study of medical and health queries to web search engines. Health Information & Libraries Journal. 2004;21(1):44-51.
16. Yang CC, Winston F, Zarro MA, Kassam-Adams N. A study of user queries leading to a health information website: AfterTheInjury. org. Proceedings of the 2011 iConference: ACM; 2011:267-272.
17. White RW, Horvitz E. From health search to healthcare: explorations of intention and utilization via query logs and user surveys. Journal of the American Medical Informatics Association. 2013.
18. Jadhav A, Andrews D, et al. Comparative Analysis of Online Health Queries Originating From Personal Computers and Smart Devices on a Consumer Health Information Portal J Med Internet Res 2014;16(7):e160
19. White RW, Horvitz E. Cyberchondria: studies of the escalation of medical concerns in web search. ACM Transactions on Information Systems (TOIS). 2009;27(4):23.
20. Winkleby MA et al. Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. American journal of public health. 1992;82(6):816-820.
21. Celler BG, Lovell NH, Basilakis J. Using information technology to improve the management of chronic disease. Medical Journal of Australia. 2003;179(5):242-246.
22. Lorig KR, Ritter PL, Laurent DD, Plant K. Internet-based chronic disease self-management: a randomized trial. Medical care. 2006;44(11):964-971.
23. Zyzzyva: The Last Word in Word Study http://www.zyzzyva.net/wordlists.shtml Accessed March 9, 2014.
24. Hunspell dictonary. http://hunspell.sourceforge.net/. Accessed March 9, 2014.
25. OpenMedSpel for Hunspell http://www.e-medtools.com/Hunspel_openmedspel.html Accessed March 9, 2014.
26. Toutanova K, Klein D, Manning CD, Singer Y. Feature-rich part-of-speech tagging with a cyclic dependency network. Paper presented at: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 12003.
27. Spink A, Wolfram D, Jansen MB, Saracevic T. Searching the web: The public and their queries. Journal of the American society for information science and technology. 2001;52(3):226-234.