Online Information Seeking for Cardiovascular Diseases: A Case Study from Mayo Clinic

Ashutosh Jadhav[a,1], Stephen Wu[b], Amit Sheth[a] and Jyotishman Pathak[b]
[a]*Knoesis Center, Wright State University, Dayton, OH, USA*
[b]*Mayo Clinic, Rochester, MN, USA*

Abstract. The objective of this study is to understand the types of health information (health topics) that users search online for Cardiovascular Diseases, by performing categorization of health search queries (from Mayoclinic.com) using UMLS Metamap based on UMLS concepts and semantic types.

Keywords. Health Information Seeking, Search Log analysis, UMLS Metamap

Introduction

Since early 2000, Internet literacy and the number of Internet users have increased significantly, including using the Internet for health information[1,2]. According to the latest 2013 Pew Survey[2], one in three American adults have gone online to find out information about a medical condition. One of the most common ways to seek online health information is via Web search engines, such as Google, with approximately 8 in 10 online health inquiries originate from a search engine[2]. Since cardiovascular disease (CVD) is one of the most common chronic diseases[3], in this work we perform a categorization of CVD-related queries into selected health categories to understand the types of information related to CVD sought by Internet users.

This study provides a better understanding of what Online Health Information Seekers (OHIS) search for and gives us better insights into their health information needs. Such information can empower us with knowledge to improve the health search experience as well as to develop advanced knowledge and content delivery systems. Previous studies have used different approaches for the classification of health search queries based on diverse categorization objectives[4-6]. This study distinctively uses a semantic approach to understand "consumer oriented" health categories for CVD-related searches. With data from MayoClinic.com[7], we categorized 10 million CVD queries into 17 health categories using UMLS Metamap[8] based on UMLS concepts and semantic types. As per our study, 'Vital Signs', 'Symptom', 'Treatment', 'Drugs and Medications', 'Diet', and 'Cause' are the top health categories that users search for concerning CVD. We also identified that 80% of the CVD search queries are categorized into either one or two health categories.

[1] Corresponding Author.

1. Methods

1.1. Data Source and Dataset Creation

In this study, we collected CVD-related search queries that direct an OHIS from Web search engines to the Mayo Clinic's consumer health information portal[7], which is one of the top online consumer-centric health information portals in the US. Our recent Web analytics statistics indicate that the MayoClinic.com portal is on average visited by millions of unique visitors every day. 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 the English language and were collected between September 2011-August 2013. Our final analysis dataset consists of 10,408,921 search queries.

1.2. Data Analysis

Mapping of search queries to UMLS: We performed a semantic analysis of health search queries by mapping all the search queries from the dataset to UMLS concepts and semantic types using UMLS Metamap[8]. Metamap is a tool for recognizing UMLS concepts in text. For a given search query, Metamap identifies one or more UMLS concepts, their semantic types, concept unique identifiers (CUIs), and other details. For example, 'Symptoms of heart attack' is mapped to the 'Symptom' and 'Heart Attack' concepts, and 'SOSY' and 'DSYN' semantic types. Refer to Table 1 for semantic type abbreviations used in this paper.

Selection of health categories: There are many possible health categories of interest. In this work, based on empirical evidence we selected 17 health categories **(Table 1)** that are more "consumer oriented" as well as can reveal details about what OHISs generally search for in the context of CVD. While selecting the health categories, we also considered 1) the health categories on popular health websites (Mayo Clinic, WebMD, etc.), 2) the types of information mentioned along with CVD search queries such as vital signs (blood pressure, heart rate), quantitative measurements (blood pressure readings, medication dosage), age groups (infants, adult, elder), and 3) the most frequent semantic types that surfaced from a semantic analysis of the CVD query log dataset using UMLS Metamap. Note that there can be possible overlaps between some health categories, for example 'Drugs & Medications' can be considered as part of 'Treatment', but in our analysis we considered both as separate health categories in order to study search traffic for each topic separately.

Categorization Approach: After mapping CVD queries to UMLS, we categorized the search queries into 17 different health categories as following:

1) UMLS has 140 semantic types and some of them 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 ('Test & Diagnosis') we utilized multiple semantic types ('DIAP','LBPR','LBTR'). In this case, we categorized all the search queries

with at least one semantic type ('DIAP', 'LBPR', 'LBTR') into 'Test & Diagnosis' category.

3) For a few health categories ('Food & Diet'), certain concepts that are closely associated with the health topic are not mapped to the selected semantic type. In such cases, we utilized both semantic types and well as semantic concepts for categorization. For example, the 'FOOD' semantic type does not include concepts such as 'meal', 'menu', 'diet', 'recipe', 'lunch', etc. as they are not actually food items. We categorized all the search queries that have a 'FOOD' semantic type or at least one concept ('menu', 'diet', 'recipe', etc.) into a 'Food & Diet' category.
4) For a few health categories (e.g., 'Cause') there are no directly associated semantic types. In such cases, we utilized semantic concepts associated with search queries. For example, we categorized all the search queries which have either 'Cause' or 'Reason' semantic concepts into the 'Cause' category.
5) For 'Living with' and 'Side effect' health categories, apart from semantic concepts, we also considered the presence of keywords ('Living with' and 'Side effect') within the search query as 'Living with' is not a concept in UMLS and we found that few search queries with side effects are not mapped to 'Side effect' concept.

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. While selecting semantic types and concepts for a health category we considered their frequency in the dataset and their scope. We manually evaluated search queries in each categories and added/removed some semantic types and concepts. We performed several iterations evaluating the semantic type/concepts for each category and we defined the categorization scheme **(Table 1)**.

Table 1. List of health categories and their respective UMLS semantic types/concepts used categorization. **Abbreviations**: SOSY-Signs and Symptoms, ORCH-Organic Chemical, PHSU- Pharmacologic Substance, 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, QNCO-Quantitative Concept, AGGP-Age Group, TMCO-Temporal Concept, Ref- Reference

Health categories	UMLS Semantic Types (**ST**), UMLS Concepts (**CC**) and Keywords (**KW**)	**Examples Search Queries**
Symptom	**ST**: SOSY **CC:** symptoms, signs, Heart murmur	Stroke symptoms
Cause	**CC**: cause, reason	Stroke cause
Risk-Complication	**CC**: risk, complications	Stroke complications
Drugs and Medications	**ST**: ORCH|PHSU, CLND, PHSU **CC**: medication, medicine, drugs, dose, dosage, remedy, tablet, pill	aspirin, Tylenol, blood pressure medication
Treatment	**ST**: TOPP, FTCN(treat*, surgery), CNCE(treatment)	Stroke treatments
Test & Diagnosis	**ST**: DIAP, LBPR, LBTR	Echocardiogram
Food and Diet	**ST**: FOOD CC**:** caffeine, recipe, meal, menu, diet, eat, breakfast, lunch, dinner, alcohol	heart healthy recipes, alcohol blood pressure
Living with	**CC:** control, manage, reduce, lower, coping, survive, survival, cure **KW:** living with	Living with high blood pressure
Prevention	**CC:** prevent, Avoidance, low risk	Stroke prevention
Side effect	**CC:** side effect **KW:** side effect	blood pressure medicine side effects
Medical device	**ST**: MEDD	Pacemaker risks
Diseases	**ST**: DSYN, **CC:** arrhythmia, avascular necrosis	Heart failure, Stroke
Quantitative	**ST**: QNCO	blood pressure ranges
Age-group Ref.	**ST**: AGGP	hypertension in children
Body part Ref.	**ST**: BPOC, BLOR	heart ablation
Vital signs	**CC:** blood pressure, heart rate, pulse rate, temperature	Blood pressure
Temporal Ref.	**ST**: TMCO	Morning high blood pressure

2. Results

Table 2. Categorization of CVD search queries into 17 health categories and their percent distribution

Sr. No	Health categories	Total Queries	% Distri-bution	Sr. No	Health categories	Total Queries	% Distri-bution
1	Vital signs	4,824,220	29.09	10	Test & Diagnosis	538,387	3.25
2	Diseases	3,436,391	20.72	11	Quantitative	377,475	2.28
3	Symptom	1,423,663	8.58	12	Living with	370,645	2.23
4	Body parts Reference	1,151,465	6.94	13	Risks and Complications	277,294	1.67
5	Treatment	1,008,587	6.08	14	Temporal Reference	184,055	1.11
6	Drugs and Medications	739,177	4.46	15	Prevention	136,428	0.82
7	Food and Diet	739,045	4.46	16	Age-group Reference	87,929	0.53
8	Medical device	665,484	4.01	17	Side effect	25,655	0.15
9	Cause	599,895	3.62		**Total**	16,585,795	100

Based on **Table 2**, the most popular health category while searching for CVD information is 'Vital signs', which includes search queries related to blood pressure, heart rate, etc. One in every five searches explicitly mentions at least one disease in the search query (such as stroke, heart failure). Other popular health categories that users search for included 'Symptom', 'Treatment', 'Drugs and Medications', 'Diet', and 'Cause'. We observe that many CVD related search queries have reference to body parts (left arm, stomach, heart), medical devices (pacemaker), quantitative measures (blood pressure readings, medication dosage), temporal reference (night, morning), and age-group (infant, child, adult, elders). Although CVD can be prevented with some lifestyle and diet changes, interestingly very few OHISs search for CVD 'Prevention'.

Table 3. A search query can be categorized into one or more health categories. The table shows the distribution of search queries by the number of health categories in which they are categorized.

Number of health categories	Number of search queries	Percentage Distribution		Number of health categories	Number of search queries	Percentage Distribution
0	615,186	5.91		5	20,313	0.20
1	4,746,969	45.60		6	4,376	0.04
2	3,561,248	34.21		7	39	0.0003
3	1,254,924	12.06		**Total**	10,408,921	100
4	205,866	1.98				

Using our categorization approach, we categorized 94% of the search queries, out of 10 million CVD related queries, into at least one health category (Table 3). Most of the queries are categorized into either one (45%) or two (34%) categories (Table 3). Very few CVD queries are categorized into 4 or more categories (2%). Our approach did not categorize 6% 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 17 categories for example, cardiac surgeon, cardiology mayo, video on cardiovascular, pediatric cardiology, and orthostatic.

3. Discussion

In this study, we categorized 94% of 10 million CVD related search queries into 17 health categories using UMLS Metamap and based on UMLS concepts and semantic types of the queries. We found using Metamap and UMLS concepts/semantic types is a very good approach for categorization of health related search queries as UMLS incorporates a variety of medical vocabularies and concepts, and mapping of each concept to semantic types. For example, using UMLS we can annotate text with a large number of drugs such as aspirin and Tylenol. However one disadvantage in using some UMLS semantic types for categorization is that some undesired concepts (in the context of our customized categorization, not in terms of UMLS concept hierarchy) are included in the semantic type. For example, semantic types 'ORCH/PHSU' and 'ORCH' are associated with the 'Drugs and Medication' category. These semantic types include some concepts that are not considered drugs to a consumer/lay population: caffeine, alcohol, fruit, prevent, etc. At the same time, if we do not consider 'ORCH/PHSU' and 'ORCH' in the drug category then we miss out on important drugs such as aspirin, and Tylenol. Online health information plays a vital role in improving health literacy and help OHIS to make more informed health decisions. Therefore, it is crucial to understand what health topics an online user may search for, as it gives us better understanding about their information needs, which can be utilized to alleviate health information search processes. This study identifies frequently searched health categories for CVD and demonstrates utility of UMLS MetaMap, semantic types and concepts for customized categorizations. The study extends our knowledge about online health information seeking and information needs in chronic diseases and particularly in CVD. Such knowledge can be used to improve Web search engines and do a better organization of health information content on health websites/applications.

Acknowledgement:

This project was supported by the Mayo Clinic and by Grant Number UL1 TR000135 from the National Center for Advancing Translational Sciences (NCATS).

References

[1] Tustin N. The role of patient satisfaction in online health information seeking. *Journal of Health Communication.* 2010;15(1):3-17.

[2] Fox S, Duggan M. Health online 2013. *Pew internet & American Life Project* 2013.

[3] National Center for Health Statistics. Health, United States, 2010: With special feature on death and dying. Hyattsville, MD. 2011.

[4] Cartright M-A, White RW, Horvitz E. Intentions and attention in exploratory health search. *SIGIR* 2011.

[5] 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.

[6] Dogan RI, Murray GC, Névéol A, Lu Z. Understanding PubMed® user search behavior through log analysis. *Database: the journal of biological databases and curation.* 2009.

[7] Mayo Clinic health information portal http://www.mayoclinic.com/ (Accessed on Feb 3, 2014)

[8] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium* 2001.