External Information Sharing on Health Forums: An Exploration

Dana M. Nguyen,[1] **Alexandra Olteanu,**[2] **Emre Kıcıman**[2]
[1]University of California, Santa Barbara
[2]Microsoft Research
dananguyen@ucsb.edu, {alexandra.olteanu, emrek}@microsoft.com

Abstract

Online health forums are an important avenue for receiving support and learning about fellow patients' experiences with similar diagnoses. We seek to characterize the external information shared (via web links) on health forums as a proxy to participants' information needs. For this purpose, using a dataset of web links shared publicly on a lung cancer forum over a period of 16 years, we perform a comparative analysis with three different website typologies, uncovering a diverse ecosystem of websites. We also examine typological variations as this forum gains and then loses popularity over time.

Introduction

Over the past decades, patients and caregivers have turned to a diversity of online resources for support and information about their or their loved one's health issues. Such resources range from health and well-being news and information websites (like WebMD) to a rich ecosystem of health communities formed on dedicated forums or social media platforms (like csn.cancer.org hosted by the American Cancer Society or r/Health/ on Reddit). A 2013 Pew Research study reported that about 7 in 10 US internet users searched for health topics online (Fox and Duggan 2013)– up from 38% in a '98 Harris Poll (The Harris Poll 2011). Their searches included diverse topics such as medical diseases and diagnoses, treatments and procedures, prescription drugs, and health insurance. Patients have also taken on a more active role by striving to produce and maintain reliable online resources that include "the best online links, the best medical centres, the best treatments, and the latest research" and cover "topics that clinicians may consider secondary but are very important to the patient–their quality of life, the impact of their disease on their friends and family, and the psychological aspects of their illness" (Ferguson 1996).

Within the ecosystem of online health communities, health forums remain a critical avenue to directly seek out and exchange support, information, and stories about health conditions. Patients often seek health information online as a key complement to information received from their doctors (Zhao and Zhang 2017; Ofran et al. 2012), using social platforms to satisfy different kinds of information needs than they do via web search (Morris, Teevan, and Panovich 2010). While the information shared on health forums is centered around personal experiences and word-of-mouth information from friends and family (Bond and Ahmed 2016; Zhao and Zhang 2017), users also bring in information from external sources via web links to external websites.

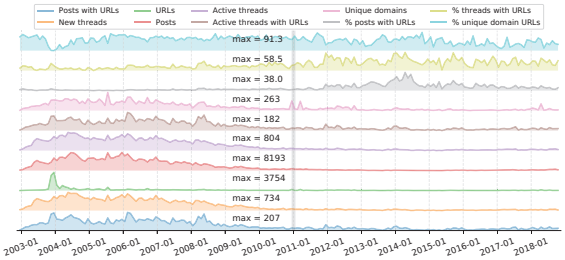

Figure 1: Normalized temporal patterns in information sharing on a health forum over a 16 years period. See in color.

The process of sharing and discussing information on health forums is akin to collective sensemaking, wherein participants—by posting, questioning, iterating—construct a shared knowledge structure (Mamykina, Nakikj, and Elhadad 2015). Under this framing, suggesting external resources is a key step in the process of collective sensemaking. Our descriptive study examines what *external resources* are being suggested in a health forum over time, and how that is changing. This provides context for research on the information needs of health patients, data mining of health forums and social media activities, and other work that contextually interprets the activities of health communities.

Research Scope. We explore what information needs are satisfied in an online health forum, and how this changes as the forum's popularity and traffic increases, plateaus and then declines over the course of about 16 years. We analyze variations in the kinds of information being shared as the forum transitions from a popular social forum to a low-trafficked forum. Does the content on the site change to reflect the changes in its nature? Or do we see the kinds of resources that are shared stagnate as the remaining partic-

	Total	P	<2011	>2011
Posts	369,284	-	344,257	25,027
Threads	38,209	-	34,451	3,758
URLs	28,302	7.7	24,035 (P = 7)	4,267 (P = 17)
Posts with URLs	11,112	3	8,907 (P = 2.6)	2,205 (P = 8.8)
Threads with URLs	9,490	24.8	7,548 (P = 21.9)	1,942 (P = 51.7)
Unique URLs	11,825*	-	9,141	2,876
Unique Domains	4,645*	-	3,841	1,135

Table 1: Dataset figures, including for the first vs. last 8 years. P – average prevalence of URLs (resp. posts/threads with URLs) per 100 posts (resp. threads). * indicates estimated values due to e.g., expired shortened URLs.

ipants maintain their static, learned behaviors, without the dynamicism created by new members (Danescu-Niculescu-Mizil et al. 2013)? We are interested in two questions (RQs):

R1) What types of information needs are met through web links sharing? To characterize the information shared via web links, we experiment with multiple categorization paradigms offering insights at different levels of granularity and partitioning rationale. While comparative analyses of multiple typologies can be illuminating (Burnett and Buerkle 2004), they remain rare.

R2) How has the information shared via web links changed over time? We conjecture that broad variations in the type of information shared on health forums across years can offer insights into how the role of these forums evolves over time.

Data Collection and Annotation

Exploring changes in the type of external content shared on health forums requires a longitudinal dataset of web links shared via health forum messages. To build such a dataset, we use public messages shared on a forum dedicated to lung cancer and maintained by a leading lung cancer nonprofit, covering about 16 years from Jan. 2003 to Nov. 2018. From these messages we then extracted shared web links (URLs),[1] and removed user metadata and post content.[2] For each URL (e.g., https://med.nyu.edu/research/office-science-research/clinical-research/), we extracted the domain (e.g., nyu.edu), which we used for categorization and analysis. Overall, URLs were shared in 25% of threads and 3% of posts. Table 1 includes our dataset figures.

Qualitative Exploration. To characterize the websites being shared, we started with a qualitative review of small samples of 25-50 URLs, containing either full URLs or their domains. We see patterns emerging from the open coding of the full URLs: a majority of webpages are health related, with a sizable fraction linking back to forum posts or to other links on the hosting NGO website. Among these are self-help websites (health information), research articles (medical research) and coverage of such articles (health news), drug manufacturing, hospital or other health service providers' websites (health care services), as well as links

[1]We used the *href* attribute that specifies the destination of a link in HTML to extract URLs, and discarded malformed URLs e.g., pointers to uploaded media and emails.

[2]Our study is IRB approved, and the data used in our analysis was stripped of any PII; this study is concerned with only the information shared on health forums via web links.

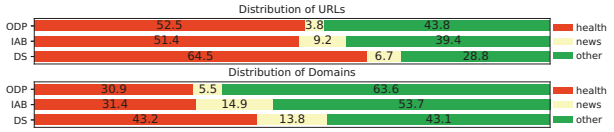

Figure 2: Prevalence of health related URLs and domains.

to other health forums. When coding at domain level,[3] other types of sites also emerge including personal blogs, diaries, or memorial sites, as well as fundraising websites. Links to NGOs, mainstream news, and governmental websites were also present in these samples. Drawing from these qualitative insights and our focus on the types of information shared via web links, we center our analysis on the URL domains.

Websites Coding

To examine the information shared via web links, we juxtapose multiple website typologies, including a domain typology drawn from prior studies of online health communities and our qualitative exploration and two existing open-domain and general purpose typologies of online content. This also ensures that our observations are not an artifact of a certain typology. Throughout the paper, for comparison purpose (when applicable) we group similar categories.

Webpage version. Given that many websites in our dataset were shared several years back, we used the Internet Archive's API (https://archive.org) to recover their archived versions. When available, we based the typological assessments on the version of the website archived closest to the time the website was initially posted onto the forum.

General purpose typologies. We classified web domains using two web typologies: 1) the Open Directory Project (ODP) classification (AOL, Inc. 2016), and 2) the Interactive Advertising Bureau (IAB) classification (IAB Tech Lab 2017). To categorize domains according to the IAB typology we used the WebShrinker API (DNSFilter, Inc. 2020), while for ODP we used an existing classifier (Bennett, Svore, and Dumais 2010). We could not categorize 12.3% of URLs with ODP, and 7.5% with IAB (in part due to outdated links) and removed them from the analysis.

Domain specific (DS) typology. In addition to these typologies—designed to be open domain and for general purpose—we also developed a typology grounded in our application domain (see Table 2). To identify the most likely category for each URL domain, we employed crowd judges to manually annotate them by selecting the most applicable category from our typology. A brief definition and examples for each category were also provided, and the judges had to pass a qualification task before being allowed to annotate. We randomly sampled 1200 domains, and for each domain we gathered 3 annotations and kept the majority label.

Exploratory Analysis & Discussion

We begin with an overview of general patterns in web links sharing. To address our guiding RQs, we then use the typolo-

[3]This merges different URLs under the same domain, helping us surface rarer domains.

Category	Description	Mentions in related work	Examples
Mainstream News Media	news media sites that cover a diversity of topics	celebrity news (Ofran et al. 2012; Quinn et al. 2013); news sources (Madden et al. 2012); news sites (Nath et al. 2016)	cnn.com, bbc.com, newswise.com
Health News & Info.	medical & health related news articles, blogs, information or education sites	technical medical information (Ferguson 1997); specialized websites (Fox and Duggan 2013); health information (Devine et al. 2016); cancer-related blogs, educational sites (Quinn et al. 2013) encyclopedic medical sites (Madden et al. 2012)	cancerindex.org, emedicine.com, webmd.com
Medical/Health Research	health and medical research articles, journals or proceedings, or research institutions	research on conditions or treatments (Ferguson 1996; De Choudhury et al. 2014; Ofran et al. 2012); .edu websites (Nath et al. 2016); academic, technical & scientific resources (Glowniak 1995; Madden et al. 2012); peer-reviewed journals (Quinn et al. 2013; Reynolds et al. 1995)	health.ucdavis.edu, nichd.nih.gov, pubmed.org, rider.edu
Health Care Services	doctors & other health care providers, pharmaceutical & other health services, products & tech	medical centers, treatments (Ferguson 1996; Glowniak 1995; Madden et al. 2012; Barney, Griffiths, and Banfield 2011); commercial, health facility (Quinn et al. 2013); health products & services (Devine et al. 2016)	spectrumhealth.org, pparx.org, novarx.com, cduma.com
Health Forums	forums and online messaging boards focused on health related issues	support groups & self-help communities (Ferguson 1997); online health or patient communities and forums (Nath et al. 2016; Elhadad et al. 2014; Kanthawala et al. 2016; Huh, Patel, and Pratt 2012; Mao et al. 2013); health discussion forums (Quinn et al. 2013; Bond and Ahmed 2016); medical message boards (Benton et al. 2011, Barney et al. 2011)	cancergrace.com, healthboards.com, forums.cancerhealth.com
Fundraising & NGOs	non-profit organizations, fundraising & awareness campaigns (by both for- & non-profit entities)	for-profit organizations (Nath et al. 2016; Madden et al. 2012); medical crowdfunding (Kim et al. 2017); non-profit (Madden et al. 2012; Nath et al. 2016); charity (Quinn et al. 2013); support organizations, awareness, and novelty items (Ofran et al. 2012)	speciallove.org, jimmyfund.org
Governmental Sources	governmental websites & information, not necessarily on health	.gov websites (Nath et al. 2016); government-initiated webpage (Kanthawala et al. 2016); governmental agencies (Madden et al. 2012; Quinn et al. 2013)	ssa.gov, statistics.gov.uk
Social Media	any type of social media sites	social media (Mao et al. 2013; Griffis et al. 2014; Ofran et al. 2012); social network (Fox and Duggan 2013; Quinn et al. 2013); discussion boards, social bookmarking (Koskan et al. 2014)	twitter.com, facebook.com, reddit.com
Personal Sites	personal sites like blogs, personal diaries, memorial websites	personal websites (Kanthawala et al. 2016; Nath et al. 2016); individually run websites (Quinn et al. 2013); personal patient blogs (Gualtieri and Akhtar 2013)	bonobike.blogspot.com, griefhealing.com

Table 2: The domain typology of websites (with examples) and the relationship to aspects mentioned in prior work.

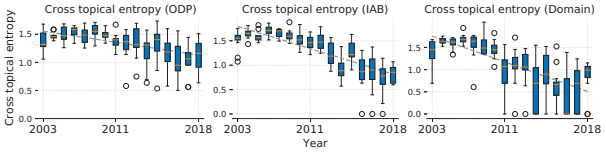

Figure 3: Variations in cross topical entropy over time.

gies to describe the web resources shared on the forum (R1) and measure topical variations through its lifespan (R2).

General trends. We observe an overall decrease in activity and an increase in the prevalence of forum threads and posts containing web links around 2009 to 2011 (Fig. 1). In fact, URLs were shared at a 2–3 times higher rate in the last vs. the first 8 years; e.g., with 8.8% vs. 2.6% of posts and 51.7% vs. 21.9% of threads including URLs (Table 1). Studies of the growth and decline of online platforms have hypothesized potential mechanisms by which platforms become less popular over time (Kairam et al. 2012; Ribeiro 2014; Durant et al. 2010), including insufficient degree of word-of-mouth growth, the community not adapting to changing environments, or a saturation of content as, e.g., once shared the websites remain a permanent informational resource.

Web resources type & prevalence (R1). Ferguson describes how patients establish "*self-help communities in cyberspace*" devoted to various health topics, like certain types of cancer, to provide "*technical medical information, practical coping tips, emotional support, and online second opinions*," with about half of the forum posts being medical in nature (Bond and Ahmed 2016; Quinn et al. 2013). Our findings concur, showing that health websites dominate the URLs shared on the forum (51.4%–64.5%), irrespective of the typology (Fig. 2). Of these, about 6% are self-references to the forum's own domain. On average, health resources were also more frequently shared, at a 1.5 (with DS typology) to 1.8 (with IAB typology) higher rate than the average web domain in our set.

Domain specific vs. general purpose typologies. Health forum users reference a diversity of topics, like medications, symptoms and treatments (Elhadad et al. 2014; Mao et al. 2013). We find this diversity also reflected in the knowledge shared via web links, both when looking at health and non-health topics. While on average, about 43% of domains (corresponding to 65% of URLs) were assessed as health related, they spanned a variety of resources such as health news and information, medical research, health care services, and references to other health forums. Among the remaining resources, 13.8% were identified as mainstream news, 6.1% as governmental sites, fundraising and NGOs, 5.8% as social media and personal sites; and the last 31.3% as miscellaneous. By aggregating health categories in the ODP and IAB typologies, we found fewer health related domains: about 31% of domains (corresponding to 52% of URLs). The remaining web domains included 17% (IAB) / 28% (ODP) focused on lifestyle, personal interests, and society; 9% (ODP) / 19% (IAB) featuring news and science content, with other secondary topics including business and technology.

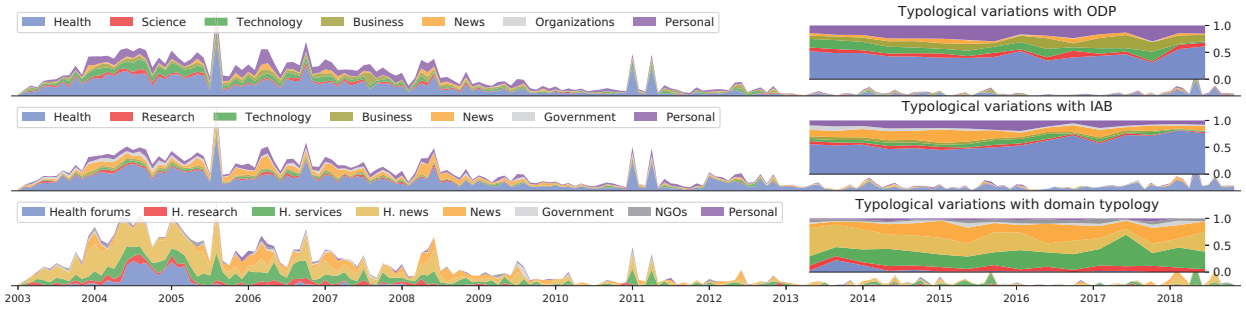

Figure 4: Topical trends over time across the three typologies. The inset plots show the normalized distribution. See in color.

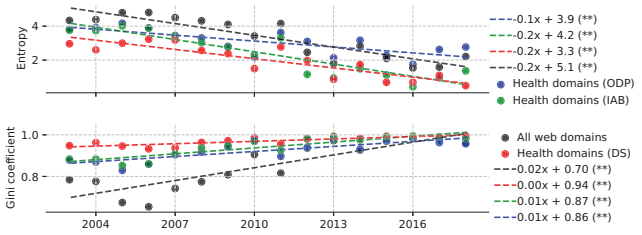

Figure 5: Yearly entropy & Gini coef. for the number of posts per web domains. $p < 0.01$ (**). See in color.

Implications & prior work. Patients turn to health forums to share their own narratives, ask questions about treatment and survival (Eschler et al. 2015), and construct and make sense of a shared knowledge structure (Mamykina et al. 2015). Web knowledge is often shared as a form of signposting and combined with users' own experiences (Bond and Ahmed 2016) to spread information, extend support to others and complement their own evidence, as well as an alternative form of communication (Nath et al. 2016). Users of online health communities also take distinct roles that exhibit different needs, sharing behaviors and motivations, such as opportunists, scientists, adventurers, and caretakers (Huh et al. 2016). Indeed, our DS typology helped uncover an interplay between different sorts of health content like news, services, and research that covers a variety of health informational needs (Fig. 4), while the more general typologies also stress the importance of more socio-emotional needs surfacing resources about hobbies and more personal activities (Fig. 4).

Typological variations over time (R2). By analyzing the entropy of the typological distribution (Fig. 3), we observe that as the forum loses popularity, the entropy also decreases—consistent with prior work showing a convergence of norms and conventions, e.g. (Danescu-Niculescu-Mizil et al. 2013), particularly as a community shrinks. Inspecting the distribution of topical interests as the popularity of the forum declines over time, we see that the core focus of the remaining participants is on health (IAB) topics, while ancillary topics become less prominent among shared web links. We hypothesize that the decrease in cross topical entropy could, in part, be a result of a convergence toward a smaller set of external resources that users turn to. When measuring the dispersion of shared domains [4] —i.e., are many domains shared, or are mostly a small number of domains shared—using two different metrics, entropy and Gini coefficient of inequality (Fig. 5), we see that there is a significant decrease in entropy and increase in Gini over time. This indicates that, indeed, a smaller set of external resources are collecting a greater degree of the attention.

Implications & prior work. Changes over time can reflect shifts in both the needs of individuals and of their communities. As users get diagnosed or as the communities establish, users may be more likely to seek information about disease and treatment related information (Ofran et al. 2012). As their treatments evolve or as the shared knowledge structure matures and saturates, their focus may also shift.

Conclusions & Limitations

As social platforms became popular, so did website sharing as a way to discover information via *word-of-mouth* (Rodrigues et al. 2011). The history of self-help websites and health communities, and of the efforts to understand these communities and their experiences goes over two decades back (Rice and Katz 2000; Ferguson 1997; Reynolds, Sharma, and Jack 1995; Ferguson 1996). Though empirical and descriptive in nature, our longitudinal study extends this literature by analyzing patterns in external resource sharing over the course of 16 years on a cancer forum, showing that such online communities appear to maintain a focus on health information, even as the forum loses popularity and as we see a drop in the diversity of content being shared.

Limitations & future work. While the scope of our study is limited by ethical considerations regarding re-use of sensitive, publicly shared content for research purposes, our findings motivate future investigations of broader scope. We observe news is shared more than government resources (Fig. 4), and the degree of attention to domains becomes unequal over time (Fig. 5): Does this behavior also appear in other websites or the web in general? Could the inequality be a side effect of search algorithms promoting sites that are already popular? Our analysis also does not capture variations of shared content (and its' heterogeneity over time) across distinct health forums and offers limited cues about factors that cause temporal variations. Future studies may

[4] To avoid a skew in the Gini coefficient from infrequently shared domains, we include only web domains posted in ≥ 4 years.

juxtapose patterns of web link sharing across multiple forums via deeper content, contextual and behavioral analyses.

References

AOL, Inc. 2016. Open Directory Project. https://dmoz-odp.org/.

Barney, L.; Griffiths, K.; and Banfield, M. 2011. Explicit and implicit information needs of people with depression: A qualitative investigation of problems reported on an online depression support forum. *BMC psychiatry* 11:88.

Bennett, P.; Svore, K.; and Dumais, S. 2010. Classification-enhanced ranking. In *WWW '10*.

Benton, A.; Ungar, L. H.; Hill, S.; Hennessy, S.; Mao, J.; Chung, A.; Leonard, C. E.; and Holmes, J. H. 2011. Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *J. of Biomedical Infor.*

Bond, C. S., and Ahmed, O. H. 2016. Can I help you? Information sharing in online discussion forums by people living with a long-term condition. *J. of Innovation in Health Informatics* 23(3):620–626.

Burnett, G., and Buerkle, H. 2004. Information exchange in virtual communities: A comparative study. *J. of Computer-Mediated Comm.* 9(2).

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *WWW'13*.

De Choudhury, M.; Morris, M. R.; and White, R. W. 2014. Seeking and sharing health information online: comparing search engines and social media. In *CHI'14*.

Devine, T.; Broderick, J.; Harris, L. M.; Wu, H.; and Hilfiker, S. W. 2016. Making quality health websites a national public health priority: Toward quality standards. *J. Med. Internet Res.*

DNSFilter, Inc. 2020. Webshrinker. https://www.webshrinker.com/.

Durant, K. T.; McCray, A. T.; and Safran, C. 2010. Modeling the temporal evolution of an online cancer forum. In *Health Informatics Symposium*.

Elhadad, N.; Zhang, S.; Driscoll, P.; and Brody, S. 2014. Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions. In *AMIA Annual Symposium*, volume 2014, 516.

Eschler, J.; Dehlawi, Z.; and Pratt, W. 2015. Self-characterized illness phase and information needs of participants in an online cancer forum. In *ICWSM'15*.

Ferguson, T. 1996. Health online: How to find health information, support groups and self-help communities in cyberspace.

Ferguson, T. 1997. Health care in cyberspace: patients lead a revolution. *The Futurist* 31(6):29.

Fox, S., and Duggan, M. 2013. Health online. *Health* 1–55.

Glowniak, J. V. 1995. Medical resources on the internet. *Annals of Internal Medicine* 123(2):123–131.

Griffis, H. M.; Kilaru, A. S.; Werner, R. M.; Asch, D. A.; Hershey, J. C.; Hill, S.; Ha, Y. P.; Sellers, A.; Mahoney, K.; and Merchant, R. M. 2014. Use of social media across us hospitals: Descriptive analysis of adoption and utilization. In *J. Med. Internet Res.*

Gualtieri, L., and Akhtar, F. Y. 2013. Cancer patient blogs: How patients, clinicians, and researchers learn from rich narratives of illness. In *ITI'13*.

Huh, J.; Kwon, B. C.; Kim, S.; Lee, S.; Choo, J.; Kim, J.; Choi, M.; and Yi, J. S. 2016. Personas in online health communities. *J. of Biomedical Infor.* 63.

Huh, J.; Patel, R.; and Pratt, W. 2012. Tackling dilemmas in supporting' the whole person' in online patient communities. In *CHI'12*.

IAB Tech Lab. 2017. Taxonomy. https://www.iab.com/guidelines/taxonomy/.

Kairam, S. R.; Wang, D. J.; and Leskovec, J. 2012. The life and death of online groups: Predicting group growth and longevity. In *WSDM'12*.

Kanthawala, S.; Vermeesch, A.; Given, B.; and Huh, J. 2016. Answers to health questions: Internet search results versus online health community responses. *J. Med. Internet Res.* 18(4).

Kim, J. G.; Vaccaro, K.; Karahalios, K.; and Hong, H. 2017. Not by money alone: Social support opportunities in medical crowdfunding campaigns. In *CSCW'17*.

Koskan, A.; Klasko, L.; Davis, S. N.; Gwede, C. K.; Wells, K. J.; Kumar, A.; Lopez, N.; and Meade, C. D. 2014. Use and taxonomy of social media in cancer-related research: a systematic review. *American journal of public health* 104(7).

Madden, K.; Nan, X.; Briones, R.; and Waks, L. 2012. Sorting through search results: a content analysis of hpv vaccine information online. *Vaccine*.

Mamykina, L.; Nakikj, D.; and Elhadad, N. 2015. Collective sensemaking in online health forums. In *CHI'15*.

Mao, J. J.; Chung, A.; Benton, A.; Hill, S.; Ungar, L. H.; Leonard, C. E.; Hennessy, S.; and Holmes, J. H. 2013. Online discussion of drug side effects and discontinuation among breast cancer survivors. *Pharmacoepidemiology and drug safety*.

Morris, M. R.; Teevan, J.; and Panovich, K. 2010. A comparison of information seeking using search engines and social networks. In *ICWSM'10*.

Nath, C.; Huh, J.; Adupa, A.; and Jonnalagadda, S. R. 2016. Website sharing in online health communities: A descriptive analysis. *J. Med. Internet Res.*

Ofran, Y.; Paltiel, O.; Pelleg, D.; Rowe, J. M.; and Yom-Tov, E. 2012. Patterns of information-seeking for cancer on the internet: an analysis of real world data. *PloS one* 7(9).

Quinn, E. M.; Corrigan, M. A.; McHugh, S. M.; Murphy, D.; O'Mullane, J.; Hill, A. D.; and Redmond, H. P. 2013. Who's talking about breast cancer? analysis of daily breast cancer posts on the internet. *The Breast* 22(1):24–27.

Reynolds, T. M.; Sharma, R.; and Jack, D. 1995. Popular medical information on internet. *Lancet* 346(8969):250.

Ribeiro, B. 2014. Modeling and predicting the growth and death of membership-based websites. In *WWW'14*.

Rice, R. E., and Katz, J. E. 2000. *The Internet and health communication: Experiences and expectations*. Sage Publications.

Rodrigues, T.; Benevenuto, F.; Cha, M.; and Gummadi, K.and Almeida, V. 2011. On word-of-mouth based discovery of the web. In *IMC'11*.

The Harris Poll. 2011. The Growing Influence and Use Of Health Care Information Obtained Online. https://theharrispoll.com/the-influence-of-the-internet-on-health-care-and-the-practice-of-medicine-continues-to-increase-in-the-late-1990s-harris-first-used-the-word-cyberchondriacs-to-describe-the-people-who-go-online-for-h/.

Zhao, Y., and Zhang, J. 2017. Consumer health information seeking in social media: a literature review. *Health Information & Libraries J.* 34(4).