Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)

"Is This an STD? Please Help!" Online Information Seeking for Sexually Transmitted Diseases on Reddit

Alicia L. Nobles,[†] Caitlin N. Dreisbach,[‡§] Jessica Keim-Malpass,[‡] Laura E. Barnes[†]
Department of Systems and Information Engineering,[†] School of Nursing,[‡] Data Science Institute[§]
University of Virginia
{aln2dh, cnd2y, jlk2t, lb3dp}@virginia.edu

Abstract

Increasing incidence of sexually transmitted diseases (STDs) has prompted the public health and technology communities to innovate new measures to understand how individuals use Internet resources to attain relevant information, particularly for sensitive or stigmatized conditions. The purpose of this study is to examine recent health information seeking and needs of the r/STD community, a subreddit focused exclusively on STDs. We found that the majority of posts crowdsource information about intermediate, non-reportable STDs such as human papillomavirus (HPV). Crowdsourced information in this community focused on symptoms, treatment, as well as the social and emotional aspects of sexual health such as fear of misdiagnosis. From our analysis, it is clear that online communities focused on discussion of health symptoms have the ripe potential to influence information-seeking behavior and consumer action.

Introduction

Reportable sexually transmitted diseases (STDs) (chlamydia, gonorrhea, and syphilis) hit a record high of 2.1 million new diagnoses in the U.S. in 2016 (CDC 2017b). STDs are a substantial public health burden with approximately 20 million new STDs annually resulting in $16 billion of annual healthcare costs in the U.S. (CDC 2017b). With many people undiagnosed, this may be an underestimate of the incidence and cost. Beyond economic consequences, STDs can result in emotional distress and comorbidities (CDC 2017a).

The accessibility of online information is an attractive source for individuals who are seeking information about stigmatized conditions. Reddit, a user-curated discussion website that is organized into subreddits focusing on specific topics, is an increasingly popular source of information for the public (Duggan and Smith 2013). An STD-specific subreddit, r/STD, exists to "...help calm the anxiety that comes with a potential STD infection through education, awareness, and prevention techniques (Reddit 2017b)."

Online health information has far-reaching effects on public health by influencing healthcare decisions and outcomes (Sbaffi and Rowley 2017) and there is interest in using the Reddit platform to provide credible information to the public (Abbey 2016). Characterizing the needs of an audience, in this case the r/STD community, is an important step to create effective health communication. In this study, we characterize recent health information seeking and needs of r/STD, a previously unstudied, but active, online community, by evaluating the health-related content of posts.

Background and Related Work

Up to 80% of Internet users use the Internet as a source of health information (Fox 2011). Consumers can now access health information anytime, anywhere with limited effort that previously could only be obtained by consulting a healthcare professional.

What makes Reddit an attractive source of information despite a large number of credible online sources with information? Selecting a source for health information is commonly based on the perception of social risk and cognitive effort to understand content, sometimes resulting in irrational choices for information (Zhang 2014). Reddit has a number of distinguishing features including accessibility (publicly available), quality (individuals crowdsource and evaluate information, although quality of information is unknown), usability (passive viewing of content without an account), and interactivity (individuals can interact with others by creating an account). This makes Reddit an ideal source for seeking health information, especially for stigmatized topics, and an increasing number of researchers are examining health-related subreddits for stigmatized conditions (De Choudhury and De 2014; De Choudhury et al. 2016; Park, Conway, and Chen 2018).

Data

The study protocol was exempt by the University of Virginia Institutional Review Board for Social and Behavioral Sciences because of the use of publicly available data.

Reddit's official API (Reddit 2017a) was used to collect posts and associated metadata (date, author) from r/STD for a one-year period August 1, 2016 to July 31, 2017 totaling 1,802 posts created by 1,557 unique users.

Methods

Content of the posts was examined using prevalence of STDs, top unigrams, and topic modeling.

Prevalence of STDs

Regular expressions were used to find keywords related to STDs present in the posts. Prior to the keyword search, posts and keywords were stemmed to reduce words to their root form (e.g., syphilis to syphili). "AIDS" was corrected to a search for "aids" because the stemmed version "aid" produced false positives. Each post was labeled with the STD if it contained any of the associated keywords. For example, herpes simplex virus (HSV) includes posts that contained the stemmed variants of "HSV" OR "herpes" OR "cold sore." We further classified the STDs into three groupings consistent with the U.S. Centers for Disease Control and Prevention (CDC): (1) acute, reportable STDs (chlamydia, gonorrhea, syphilis), (2) intermediate, non-reportable STDs (HSV, human papillomavirus [HPV], and molluscum contagiosum), and (3) bloodborne chronic, reportable STDs (human immunodeficiency virus [HIV], acquired immunodeficiency syndrome [AIDS], and hepatitis). The groupings are not mutually exclusive (e.g., if a post mentioned chlamydia and HSV, then it was labeled as both an acute, reportable STD and intermediate, non-reportable STD).

Top Unigrams

Posts were pre-processed by removing special characters, removing common stop words, and joining frequently adjacent co-located unigrams into phrases (e.g., cold_sores). Unigrams were sorted by frequency to obtain the top 50 unigrams in posts based on term frequency-inverse document frequency (tf-idf).

Topic Modeling

Unsupervised topic modeling, specifically non-negative matrix factorization (NMF), was used to discover themes present in the posts based on tf-idf of unigrams. We iterated through models with topics ranging from 10 to 40 by increments of five, manually labeled the topics, and selected the model that produced the most coherent, representative topics. NMF was selected for modeling over latent Dirichlet allocation (LDA), arguably the most widely used topic model, because LDA produced substantially less coherent topics for this dataset. We suspect NMF produces higher topic coherence for niche content (O'Callaghan et al. 2015).

Results

Prevalence of STDs

As shown in Table 1, the majority of posts focused on the intermediate, non-reportable STDs (n = 816, 45%) followed by acute, reportable STDs (n = 250, 14%) and chronic, bloodborne STDs (n = 209; 12%). The prevalence of STDs in the posts aligns with the prevalence of STDs in the general U.S. population (CDC 2017b). That is, HPV and HSV are the most common STDs that posters mention and, of the reported notifiable STDs, the most commonly mentioned in order were chlamydia, gonorrhea, and syphilis. There were minimal posts related to trichomoniasis, lice, pelvic inflammatory disease, lymphogranuloma venereum, bacterial vaginosis, and mycoplasma genitalium (less than ten posts for each STD).

Top Unigrams

Table 2 presents the top unigrams contained in the posts. Posts containing health care-related language (test, negative, doctor) focused on testing for STDs, discussed a recently received diagnosis, or sought a second opinion. For example,

> . . . I went to the doctor the day that I had developed [a lesion] and she thought it was for sure an STD. She swabbed it for herpes, as well as [testing for HIV, syphilis, chlamydia, and gonorrhea]. They all came back negative . . .

The poster then discusses hyper-information seeking, including paranoia, multiple rounds of STD testing despite repeat negative results, and asks readers if they recommend "settling down" or additional testing.

Posts containing symptom-centric language (symptom, bump, pain, red, little, and itch) described their symptoms in great detail focusing on specific areas of their body (area, penis, skin), occasionally accompanied by an image of the symptoms. For example, "Painless bump on pubic area."

Posts that contained temporal language (time, start, day, week, month, and year) commonly described a sexual encounter and subsequent symptoms. For example,

> Do I have HIV? I lost my virginity to a girl I met online about 2 months ago. A few days after that my penis started burning . . .

Finally, many posts contained information seeking language that users employed when trying to understand their health concern. Action words such as look, think, know, notice, and worry were commonly used to describe the behaviors associated with making symptom comparisons and locating adequate condition-specific information.

Topic Modeling

Posts primarily focused on crowdsourcing information on transmission [a], symptoms, testing [b], and treatment [c] of STDs; expressing concern or fear over a potential diagnosis [d]; and seeking advice on disclosing to a partner [e]. Many posters focused on crowdsourcing a diagnosis with posts titled such as "What is this?" and provided a picture or description of their symptoms to assist commenters in diagnosis. Several posters focused on fears of being misdiagnosed by their healthcare provider [f] and were crowdsourcing a second opinion. These themes are expressed in the topic modeling. Twenty-three topics from the 25 topic model were coherent enough to identify a theme, as shown in Table 3.

> [a] Can someone explain [if] Herpes simples virus can be transmitted in this situation?

> [b] HIV RNA test at 11 days, OraQuick at 43 days, both negative. Am I in the clear?

> [c] Can you order the treatment for Chlamydia online?

> [d] . . . I spend my days googling first symptoms and I think I'm starting to become hypochondriac . . .

> [d] I am worried to the point of having anxiety attacks, please tell me what you think

STD	No. of Posts
HSV (herpes | hsv | cold sore)	556 (30.9%)
HPV (wart | papillomavirus | hpv | cancer)	301 (16.7%)
Acute reportable (chlamydia | gonorrhea | syphilis)	250 (13.9%)
Intermediate non-reportable (HSV | HPV | MC)	816 (45.3%)
Chronic bloodborne (HIV | AIDS | hepatitis)	209 (11.6%)

Table 1: Prevalence of STDs in posts. MC = molluscum contagiosum.

test	119	month	49	condom	36
herpes	92	pain	49	year	36
std	81	hpv	48	infect	35
like	76	hiv	48	negative	35
sex	75	feel	48	thing	35
day	74	doctor	45	oral	35
know	73	wart	45	went	35
week	67	start	44	red	34
time	67	look	44	little	34
got	66	thank	42	itch	34
bump	63	guy	42	area	37
symptom	60	want	41	tell	33
think	56	possible	40	thought	33
help	55	girl	40	right	33
penis	54	said	39	skin	32
worry	52	chlamydia	39	say	32
notice	51	use	37		

Table 2: Term frequency-inverse document frequency of top 50 unigrams in posts.

[e] Found out I have an STI, and nervous about telling my new partner. . . . I know I should tell him, but I'm scared this will be the end of us. I am so ashamed :(

[f] Misdiagnosed? Herpes or Molluscum? . . . My doctor told me I had genital herpes after a visual inspection. No testing was done to verify I have HSV . . .

Discussion

It is imperative for researchers and public health officials to understand the cultural norms and information needs of an audience in order to create effective, tailored health communication specific to an online community. Beyond tailoring health communication, monitoring of social media may be an effective, innovative strategy to complement and modernize traditional surveillance methods that rely on individuals accessing health care or participating in studies (CDC 2017b; Paul and Dredze 2017). This information can be valuable for social marketing efforts to correct misunderstandings about transmission of STDs, combat stigmatization, encourage appropriate testing and disclosure, and supply credible resources. Furthermore, monitoring may inform public health officials of trending concerns and information about emerging STDs such as mycoplasma genitalium.

STDs carry differing experiences in symptomology, fear, uncertainty, stigma, and perceptions of risk (Meites et al. 2013). These various risk perceptions likely contribute to intention of accessing the health care systems for subsequent testing and treatment (Du et al. 2011). For those who do not have routine access, the uncertainty is likely heightened (Du et al. 2011). The ability to access 'lay health expertise' through unofficial online sources may contribute to the decision to seek care and is an area of needed future study, particularly as it relates to credibility of these sources.

Limitations

Since users of r/STD do not need to identify demographic information, the true demographic is unknown and we caution in making inferences beyond this community. We acknowledge that there may be other subreddits where sexual health is discussed. Finally, we did not account for phonetic spellings or slang terms. This could impact our analysis, however, this was not prevalent in the aforementioned analyses, so we suspect this has a negligible effect on the overall interpretation.

Conclusions

As the Institute of Medicine stated and the CDC reiterated in their most recent surveillance report, STDs remain "hidden epidemics of tremendous health and economic consequence in the United States (Institute of Medicine 1997; CDC 2017b)." Despite this, resources for STDs are dwindling requiring innovative strategies for public outreach (CDC 2017b). In this paper, we explored recent online information seeking and health needs of the r/STD community. Although the number of users in our study is limited, the affected audience is likely much larger because of passive viewing of content. Our findings indicate the prevalence of sexual health information crowdsourcing on the r/STD social media platform and highlights the potential to aid in targeted health communication. Future research needs to take into consideration the immense power of online communities in influencing health seeking behavior and wellness, especially in vulnerable populations, and temporal changes in the public's information needs.

Acknowledgements

We thank Dylan Hazlett for his valuable support as an undergraduate research assistant for this study.

References

Abbey, D. 2016. Reddit community finds health answers from reference collaborative. Retrieved August 4, 2017 from https://dspace.library.colostate.edu/bitstream/handle/10968/1791/CUHSLMCM_M451.pdf?sequence=1.

CDC. 2017a. Hiv/aids & stds. Retrieved January 5, 2017 from https://www.cdc.gov/std/hiv/default.htm.

Topic Label	Unigrams	Example Post Title
HSV	herpes, outbreak, virus, ingrown_hair, pimple	“Herpes or infected follicles?”
HPV	hpv, high_risk, wart, vaccine, cancer	”Does this look like HPV?”
Symptoms	fordyce_spots, sti, std, symptoms, blister	“Fordyce spots, ppp, or std???”
Symptoms	bumps, small, itchy, look_like, clear	“Should I be worried about these bumps?”
Crowdsourcing	help_identify, concerned, opinion, worse, idea	“Could anyone help me identify this?”
Help seeking	help, reddit, figure, doctor, scared	“Please help! Kinda scared”
HIV	hiv, risk, condom, exposure, oral_sex	“Risk of getting HIV from bleeding gums”
Chlamydia	chlamydia, treatment, tested_positive, doxycycline, azithromycin	“Can doxycycline cure chlamydia that I just got yesterday”
Genital warts	wart, tyginta, acv, treatment, look_like	“Can I share Tyginta with my boyfriend even if he hasn’t got warts yet?”
Symptoms	penis, redness, shaft, dry, worry	“Bumps on Penis, Really worried”
Warts	genital_warts, look_like, terrified, tyginta, treatment	“Genital warts? 7 months pregnant - never had an std before - terrified.”
Testing	test, results, got_tested, negative, positive	“Can the results of the OraQuick test be trusted, especially if they are negative?”
Solicited sex	worried, prostitute, changes, girlfriend, sti	“I had unprotected sex with a prostitute. Should I be worried?”
Symptoms	bump, small, ingrown_hair, pimple, doesn_hurt	“Have this bump for almost 3 months now. Should I be alarmed?”
Risk	unprotected_sex, years, partners, months, tested	“Is there a place to get tested asap?”
MC	molluscum, vinegar, cured, spread, wondering	“Molluscum? Other? Help please”
General questions	sex, unprotected_oral, safe, worry, used_condom	“Freaking out! Had sex with a stranger.”
Possible STDs	got, noticed, painful, need_help, possible_std	“Help with a possible STD.”
Crowdsourcing	ideas, months, week, noticing, years_ago	“Had for over a week now...Any ideas?”
HSV	hsv, genital, oral, testing, contract	“Contracting HSV-2 indirectly through group sex?”
Symptoms	rash, red, itchy, inner_thigh, help_identify	“Single rash on pubic area.”
Crowdsourcing	opinion, thought, tested, seeing, going	“NSFW - Your opinion on what’s going on”

Table 3: Topics present in posts. MC = molluscum contagiosum.

CDC. 2017b. Sexually Transmitted Disease Surveillance 2016. Technical report, U.S. Department of Health and Human Services, Atlanta, GA.

De Choudhury, M., and De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, ICWSM-14, 71–80.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proc SIGCHI Conf Hum Factor Comput Syst*, CHI ’16, 2098–2110. New York, NY, USA: ACM.

Du, P.; Camacho, F.; Zurlo, J.; Lengerich, E.; Du, P.; Camacho, F.; Zurlo, J.; and Lengerich, E. J. 2011. Human immunodeficiency virus testing behaviors among us adults: the roles of individual factors, legislative status, and public health resources. *Sex Transm Dis* 38(9):858–864.

Duggan, M., and Smith, A. 2013. 6% of online adults are reddit users. Technical report, Pew Research Center, Washington, D.C.

Fox, S. 2011. Health topics: 80% of internet users look for health information online. Technical report, Pew Research Center, Washington, D.C.

Institute of Medicine. 1997. *The Hidden Epidemic: Confronting Sexually Transmitted Diseases*. Washington, DC: The National Academies Press.

Meites, E.; Satterwhite, C. L.; Torrone, E.; Meites, E.; Dunne, E. F.; Mahajan, R.; Ocfemia, M. C. B.; Su, J.; Xu, F.; and Weinstock, H. 2013. Sexually transmitted infections among us women and men: Prevalence and incidence estimates, 2008. *Sex Transm Dis* 40(3):187–193.

O’Callaghan, D.; Greene, D.; Carthy, J.; and Cunningham, P. 2015. An analysis of the coherence of descriptors in topic modeling. *Expert Syst Appl* 42(13):5645–5657.

Park, A.; Conway, M.; and Chen, A. 2018. Examining thematic similarity, difference, and membership in three online mental health communities from reddit: A text mining and visualization approach. *Comput Human Behav* 78:98–112.

Paul, M. J., and Dredze, M. 2017. Social Monitoring for Public Health. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 9(5):1–183.

Reddit. 2017a. Reddit api documentation. Retrieved August 15, 2017 from https://www.reddit.com/dev/api.

Reddit. 2017b. r/std. Retrieved November 11, 2017 from https://www.reddit.com/r/STD.

Sbaffi, L., and Rowley, J. 2017. Trust and credibility in web-based health information: A review and agenda for future research. *J Med Internet Res* 19(6):e218.

Zhang, Y. 2014. Beyond quality and accessibility: Source selection in consumer health information searching. *J Assoc Inf Sci Technol* 65(5):911–927.