Crowd-Diagnosis: When the Public Turns to Social Media to Obtain Clinical Diagnoses

Alicia L. Nobles, PhD, MS,[1,2] Eric C. Leas, PhD, MPH,[2,3] Mark Dredze, PhD,[4] Christopher A. Longhurst, MD, MPH,[1] Davey Smith, MD,[1] John W. Ayers, PhD, MA[1,2]

[1]Dept of Medicine, University of California San Diego, La Jolla, CA, USA; [2]Center for Data-Driven Health, University of California San Diego, La Jolla, CA, USA; [3]Dept of Family Medicine and Public Health, University of California San Diego, La Jolla, CA, USA; [4]Dept of Computer Science, Johns Hopkins University, Baltimore, MD, USA

Introduction

In a 2019 study published in the Journal of the American Medical Association, we coined the term "crowd-diagnosis" to describe when people turn to public social media to obtain a diagnosis. Using a case study of sexually transmitted infections, we found thousands requesting crowd-diagnoses, commonly posting pictures to aid in diagnosis and sometimes seeking diagnoses to overrule a doctor's diagnosis.[1] Our goal is to extend this work to a more general setting by focusing on a popular social media forum dedicated to obtaining feedback on medical conditions and answering **(RQ1)** who requests crowd-diagnoses, **(RQ2)** for what health issues are crowd-diagnoses more frequently sought, and **(RQ3)** what crowd-diagnosis requests are most likely to receive a response?

Methods

Data. Reddit is a social media site with 330 million monthly users, primarily from the US,[2] organized into communities called subreddits. r/AskDocs is a large subreddit with 185k subscribers (January 2020) that allows users to submit questions about personal medical conditions for an opportunity to be answered by physicians that are verified by community moderators. Posters are required to provide as much detail as possible, including demographics. We collected all posts, comments, and associated metadata from its inception in July 2013 through December 2018. Our final dataset contained 190,974 posts from 131,020 posters with 694,533 comments from 106,306 commenters.

Analysis. RQ1 quantifies the self-reported demographics of posts to contextualize the users of this online community. We developed regular expressions to automatically extract a post's self-identified gender (female, male, transgender) and race/ethnicity (Asian, black, Hispanic, Indian, middle eastern, multiracial, white). The racial/ethnic categories were derived from posters' self-descriptions. **RQ2** quantifies the topic prevalence of health issues and contextualizes who (demographics) asks about these health issues. We applied a probabilistic topic model to identify the topics of the posts and compared differences in topic prevalence across demographics (e.g. what do women versus men talk about) by calculating the odds of each topic in the demographic of interest to the reference group (i.e. the majority group). **RQ3** measures the frequency of responses and the probability of a response associated with the demographics of the poster and the topic of the post. We used logistic regression to estimate the probability to estimate the probability and corresponding 95% CIs of (1) receiving any response and (2) receiving a response from a physician.

Results

RQ1. Gender was identified in 64% of all posts (test set accuracy of 100%) while race/ethnicity was identified in 33% of all posts (test set accuracy of 99%). Posts were primarily male (43% of all posts) and white (25% of all posts), although some identified as black (1.6% of all posts), Hispanic (0.8% of all posts), and transgender (0.2% of all posts).

RQ2. Crowd-diagnoses were most frequently sought for dermatological conditions (7.5%), followed by reproductive health (2.6%), interpretation of diagnostic testing (2.2%), dental and oral health (2.1%), and allergies (2.1%) (Figure 1). Crowd-diagnoses were also sought for cardiology conditions (2.1%), cancer (1.2%), and infectious diseases (1.2%), among other complex issues. Posts identifying as female or transgender sought help on sensitive topics (e.g., female-specific health including pregnancy, pregnancy loss, and menstruation) at a higher rate than their male counterparts. Females discuss chronic issues, such as fibromyalgia, at higher rates than males. Transgender people have higher odds of inquiring about hormonal issues, side effects, medication, and health care interactions than males.

RQ3. Seventy-two percent of all posts received a response. Posts that self-identified as female 72.9% (95%CI, 72.5-73.3) were more likely to receive any response than posts self-identified as male 69.6% (95%CI, 69.2-69.9). Posts

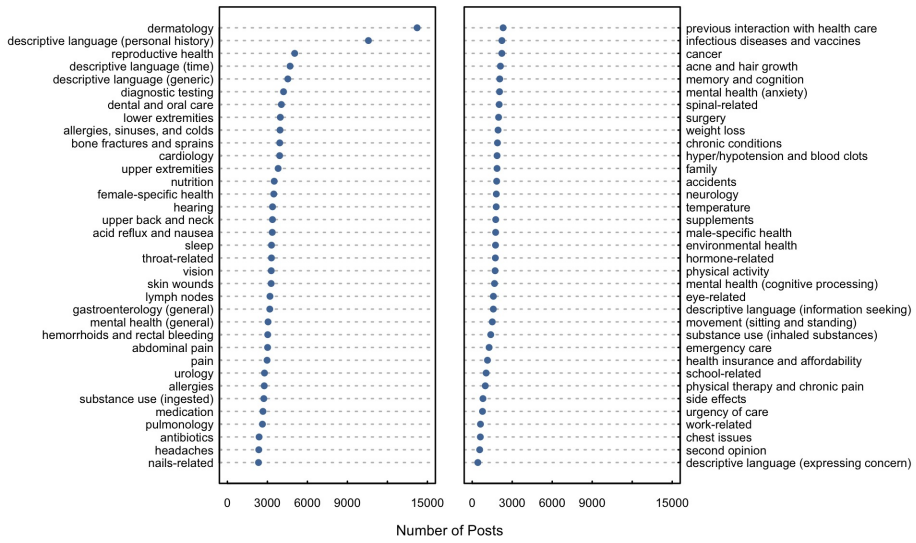

Figure 1: Dot plot of the topics present in the posts. Dot indicates the number of posts that were assigned the topic.

that self-identified as black 74.7% (95%CI, 73.2-76.2) and multi-racial 75.5% (95%CI, 73.2-76.2) were more likely to receive any response than posts self-identified as white 69.7% (95%CI, 69.3-70.2). Eleven percent of posts received a response (12% of all total comments) from a physician and the rates of receiving a response from a physician were statistically indistinguishable across the self-identified demographics in the posts.

Discussion

Extending our previous work, this study shows that crowd-diagnoses are commonly sought across health issues and often receive a reply. Posts that identify as female, black, and multi-racial were more likely to receive responses, suggesting a potential benefit for overcoming barriers to care among these populations. Although crowd-diagnoses have the benefits of anonymity, speed, and democratizing diagnoses (including those with undiagnosed illnesses)[3], they may pose potential dangers. For example, even on an online community intended to engender responses from physicians, few responses were from physicians. Further studies are needed for strategies to address this under-engagement. The public format of crowd-diagnoses could become a source of misinformation for passive viewers of a misdiagnosis, potentially resulting in a ripple effect if these viewers then wrongly self-diagnose. On the other hand, crowd-diagnoses have the potential to substantially improve public health by connecting people seeking help for medical issues in with credible resources and information. Experts could moderate requests for crowd-diagnoses, resulting in social media being a vehicle to connect the public to professional health care. Modifications, like suggested searches, can help people and moderators by ensuring similar queries use consistent language. The current norm in public health communication is to invest in top-down resources that hopefully overlaps with the needs of the public. However, few clinicians would expect so much unmet demand for remote treatment referral for this diversity of topics. Studying crowd-diagnoses broadly, beyond a singular subreddit, can identify what types of information the public is willing to share and unmet health information needs to build out evidence-based resources that are complementary.

References

1. Nobles AL, Leas EC, Althouse BM, Dredze M, Longhurst CA, Smith DM, Ayers, JW. Requests for Diagnoses of Sexually Transmitted Diseases on a Social Media Platform. JAMA. 2019;322(17):1712–1713.
2. Alexa. Competitive Analysis, Marketing Mix & Traffic. www.alexa.com/siteinfo/reddit.com. Retrieved July 2019.
3. Meyer AND, Longhurst CA, Singh H. Crowdsourcing Diagnosis for Patients With Undiagnosed Illnesses: An Evaluation of Crowdmed. JMIR. 2016;18:(1):e12