Engagement with COVID-19 Public Health Measures in the United States: A Cross-Sectional Social Media Analysis from June to November 2020

Original Paper

Daisy Massey, BA;1 Chenxi Huang, PhD;1 Yuan Lu, ScD;1,2 Alina Cohen, MS, MBA;3 Yahel Oren, MS;3 Tali Moed, BS;3 Pini Matzner, PhD;3 Shiwani Mahajan, MBBS, MHS;1,2 César Caraballo, MD;1,2 Navin Kumar, MPhil;4 Yuchen Xue, MA;5 Qinglan Ding, PhD, MBBS;6 Rachel P. Dreyer, PhD;7 Brita Roy, MD, MPH, MHS;8,9 Harlan M. Krumholz, MD, SM1,2,10

1 Center for Outcomes Research and Evaluation, Yale New Haven Hospital, New Haven, Connecticut
2 Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut
3 Signals Analytics, New York, New York and Netanya, Israel
4 Department of Sociology, Yale University, New Haven, Connecticut
5 Foundation for a Smoke-Free World, New York, New York
6 College of Health and Human Sciences, Purdue University, West Lafayette, Indiana
7 Department of Emergency Medicine, Yale School of Medicine, New Haven, Connecticut
8 Department of Medicine, Yale School of Medicine, New Haven, Connecticut
9 Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, Connecticut

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Address correspondence to:
Harlan M. Krumholz MD, SM
1 Church Street, Suite 200, New Haven, CT 06510
Telephone: 203-764-5885; Fax: 203-764-5653
Email: harlan.krumholz@yale.edu; Twitter: @hmkyale

Abstract

Background: The coronavirus disease 2019 (COVID-19) has continued to spread in the US and globally. Closely monitoring public engagement and perception of COVID-19 and preventive measures using social media data could provide important information for understanding the progress of current interventions and planning future programs.

Objective: To measure the public’s behaviors and perceptions regarding COVID-19 and its daily life effects during the recent 5 months of the pandemic.

Methods: Natural language processing (NLP) algorithms were used to identify COVID-19 related and unrelated topics in over 300 million online data sources from June 15 to November 15, 2020. Posts in the sample were geotagged, and sensitivity and specificity were both calculated to validate the classification of posts. The prevalence of discussion regarding these topics was measured over this time period and compared to daily case rates in the US.
Results: The final sample size included 9,065,733 posts, 70% of which were sourced from the US. In October and November, discussion including mentions of COVID-19 and related health behaviors did not increase as it had from June to September, despite an increase in COVID-19 daily cases in the US beginning in October. Additionally, counter to reports from March and April, discussion was more focused on daily life topics (69%), compared with COVID-19 in general (37%) and COVID-19 public health measures (20%).

Conclusions: There was a decline in COVID-19-related social media discussion sourced mainly from the US, even as COVID-19 cases in the US have increased to the highest rate since the beginning of the pandemic. Targeted public health messaging may be needed to ensure engagement in public health prevention measures until a vaccine is widely available to the public.

Keywords: COVID-19 public perception; COVID-19 social media; infodemic; social media research; social media analysis; natural language processing; Reddit data; Facebook data; COVID-19 public health measures; public health surveillance
Introduction

As the coronavirus disease 2019 (COVID-19) continues its spread in the United States (US), a key to controlling the spread until a vaccine is widely available is to enlist the public in risk-mitigation behaviors.[1,2] Studying the public’s social media posts regarding COVID-19 public health measures may provide information about targets of intervention, progress toward behavior goals, and risk of future outbreaks.[3-9] Although real-time reports on pandemic-related tests and mortality are widely available, there are fewer opportunities to gain near real-time insight into behaviors and beliefs about the pandemic.

Social media, which people are using now more than ever to communicate, has served as a useful data source in providing rapid insight into the public’s behaviors and beliefs during the pandemic.[10-13] Studies have noted a high prevalence of COVID-19-related discussion, including such topics as hygiene, shortages, and the spread of misinformation, and an increase in COVID-19-related discussion as COVID-19 cases increase.[5,14,15] However, existing findings are based on evidence during only the beginning of the outbreak, from December 2019 to April 2020, and the range of topics and keywords explored is also limited.[7,14,15-19] Additionally, studies analyzing COVID-19 behaviors and beliefs on social media have primarily used Twitter as their source, which has several limitations.[14-16] Most notably, highly rated retweets are more likely to derive from spam and bot accounts, which are also actively posting about COVID-19, and can obscure the targeting of signals from human discussions.[20-22] Further, previous studies each focused on a particular aspect of the pandemic, such as disinformation relating to the pandemic, without comparing the volume of discussion related to multiple aspects to determine the public’s relative focus on particular pandemic-related issues and behaviors.
Therefore, there is a need to assess how the public’s current reaction to the pandemic has changed since the early stages, by examining broad online discussion from more diverse sources. Accordingly, we studied changes in social media discussion of multiple topics, including daily life topics, which may be affected by the pandemic, COVID-19-related public health behavior topics, and mentions of COVID-19, from June through November 2020, and assessed their correlation with COVID-19 new daily case rates (incidence). In measuring these trends in social media data and the COVID-19 incidence rate in the US, we sought to elucidate the US public’s engagement with COVID-19-related public health measures, which are crucial to addressing the current pandemic.

Methods

Data Sources

The data sample consisted of unstructured, English-language posts from social media outlets and forums, such as Reddit, Facebook, and 4Chan and comments from news sites.[23] Signals Analytics, an advanced analytics consultant that conducted the analysis, accessed these data sources through a third-party data vendor, NetBase.[24,25] These social media posts were geotagged by NetBase both directly, by using geolocation data from posts, and indirectly, by using author profiles and unique domain codes (such as .uk). All data were deidentified by NetBase before being transferred to Signals Analytics.

In addition to the social data, the study included US COVID-19 case data from the COVID-19 Dashboard by the Center for Systems Science and Engineering at Johns Hopkins University.[26] These data were updated daily using a public application programming interface (API), and included total number of deaths, new daily deaths, total active cases, and daily new cases.[27]
No personal identifying information (e.g., usernames, emails, or IP addresses) was shared as part of the analysis or reporting process. This study was exempted from Institutional Review Board review by Yale University as it did not engage in research involving human subjects.

**Approach**

To determine trends in social media discussion during the COVID-19 pandemic, we collected data posts from all Internet sources and applied natural language processing (NLP) algorithms to identify and classify mentions of COVID-19, COVID-19-related public health measures, and daily life topics.

NetBase runs a daily query that we designed based on our project scope on over 300 million online data sources from June 15 to November 15, 2020 (eMethods 1). There were several steps to narrow the sample retrieved from the query to include only posts relevant to our research question (eTable 1). First, NLP algorithms were run to remove advertisements and pornography-related sites and posts (eMethods 2). Next, a taxonomy of topics was applied (eMethods 3). The posts that did not include discussion of topics from the taxonomy were deleted. Finally, all news articles and blog posts were deleted from the sample, so that the only remaining data posts were from social outlets (forums and comments on news sites).

The taxonomy was comprised of two categories, COVID-19-related public health measures and daily life behaviors, each of which included multiple topics (eMethods 4). COVID-19 mentions was also an individual topic in the taxonomy, independent of either category, to measure posts that directly mentioned COVID-19 by name or synonym.

Once all posts were classified according to the topics in the taxonomy, we measured trends in these topics over time by tracking the total number of posts that included mentions of each taxonomy topic and category. Classifications of topics and categories were not mutually
exclusive, so the same post was able to be classified by multiple topics across any category. Trends were visualized by taxonomy category, COVID-19 mentions, and by the most commonly mentioned taxonomy topics. These trends were visualized with the COVID-19 incidence rate in the US. This approach allowed us to identify changes in both topics that prior research in the early stage of the outbreak had shown to be prevalent in COVID-19 discussion, and in topics from daily life and COVID-19 literature reviews that were not previously known to be found in COVID-19 discussion, but may have become so or changed significantly as COVID-19 cases or current events changed.[15,16,28-32] Additionally, our approach removed redundant posts, limiting the effect of bots and reposts. The taxonomy classification was validated by calculating specificity and sensitivity (eMethods 5). In an independent data sample of 100 posts classified by manual review, the algorithm had over 80% specificity, which is higher than comparable social media research.[29] Sensitivity was calculated as the number of correct classifications of a topic using the NLP algorithms divided by the total number of posts for the topic identified by manual screening, and we found that our taxonomy approach led to an average classification rate of 92% sensitivity. We also validated the methodology by applying it to US-specific current events and found that the approach revealed an increase in online social discussion when the given current event topic was most relevant (eFigure 1). This methodology has been shown to reveal insights into outbreak characterization and event prediction for the E-cigarette or Vaping Use-Associated Lung Injury (EVALI) outbreak.[33]

**Results**

The final data sample consisted of 9,065,733 online social posts that mentioned at least one of the topics in our taxonomy from June 15 to November 15, 2020 (Table 1). The majority
(87%) of posts in our sample came from sources that were categorized as forums, such as Reddit and Facebook (Table 2; eTable 1).[23] The minority of posts (13%) in our sample were derived from comment sections on news sites, including The Hill, a media source focused on politics and business, and Breitbart, a right-leaning media source (Table 2; eTable 1).[34,35] Most posts in the sample were not able to be directly geotagged due to sources’ data privacy measures and restrictions. A minority were geotagged as from the US, with the remaining geotagged as from a country other than the US (eTable 2). Using indirect geotagging provided by NetBase, it was estimated that about 70% of all initial posts collected by the search query were from the US.

Table 1. Number of posts by taxonomy topic (June 15 to November 15, 2020)

<table>
<thead>
<tr>
<th>Relevant Taxonomy Categories</th>
<th>Relevant Taxonomy Topics</th>
<th>Number of Posts with Mentions (Percent Classified within Category)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=9,065,733</td>
<td></td>
<td>1,836,200</td>
</tr>
<tr>
<td><strong>COVID-19-Related Public Health Topics</strong> (20)</td>
<td>Wearing Face-Mask</td>
<td>1,120,344 (61)</td>
</tr>
<tr>
<td></td>
<td>Lockdown</td>
<td>457,705 (25)</td>
</tr>
<tr>
<td></td>
<td>Social Distancing</td>
<td>242,105 (13)</td>
</tr>
<tr>
<td></td>
<td>Quarantine</td>
<td>94,301 (5)</td>
</tr>
<tr>
<td></td>
<td>Testing</td>
<td>87,712 (5)</td>
</tr>
<tr>
<td></td>
<td>Excessive Hand-Wash</td>
<td>64,679 (4)</td>
</tr>
<tr>
<td></td>
<td>Contact Tracing</td>
<td>31,775 (2)</td>
</tr>
<tr>
<td></td>
<td>Reopening</td>
<td>16,681 (1)</td>
</tr>
<tr>
<td></td>
<td>Screening</td>
<td>14,569 (1)</td>
</tr>
<tr>
<td></td>
<td>Wearing Gloves</td>
<td>11,531 (1)</td>
</tr>
<tr>
<td></td>
<td>Disinfection</td>
<td>11,076 (1)</td>
</tr>
<tr>
<td></td>
<td>Wearing Face Shield</td>
<td>10,104 (1)</td>
</tr>
<tr>
<td><strong>Daily Life Taxonomy Topics</strong> (69)</td>
<td>Sex Life</td>
<td>887,457 (14)</td>
</tr>
<tr>
<td></td>
<td>Food</td>
<td>838,513 (14)</td>
</tr>
<tr>
<td></td>
<td>Financial</td>
<td>710,757 (11)</td>
</tr>
<tr>
<td></td>
<td>Travel</td>
<td>651,426 (10)</td>
</tr>
<tr>
<td>Category of COVID-19 Discussion Topic</td>
<td>Posts with Mentions of COVID-19 Related Public Health Behavior</td>
<td>Posts with Mentions of Daily Life (%) N=6,210,255</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Smoking/Vaping</td>
<td>476,468 (8)</td>
<td></td>
</tr>
<tr>
<td>Mass Gatherings</td>
<td>451,815 (7)</td>
<td></td>
</tr>
<tr>
<td>Virtual Communication</td>
<td>414,549 (7)</td>
<td></td>
</tr>
<tr>
<td>Alcohol Consumption</td>
<td>398,229 (6)</td>
<td></td>
</tr>
<tr>
<td>Religion</td>
<td>285,538 (5)</td>
<td></td>
</tr>
<tr>
<td>New Skills/Hobbies Acquisition/DIY</td>
<td>280,155 (5)</td>
<td></td>
</tr>
<tr>
<td>Drug Use</td>
<td>257,819 (4)</td>
<td></td>
</tr>
<tr>
<td>News/Media Consumption</td>
<td>257,415 (4)</td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>246,074 (4)</td>
<td></td>
</tr>
<tr>
<td>Physical Activity</td>
<td>205,116 (3)</td>
<td></td>
</tr>
<tr>
<td>Work from Home</td>
<td>198,057 (3)</td>
<td></td>
</tr>
<tr>
<td>Socializing in Person</td>
<td>177,522 (3)</td>
<td></td>
</tr>
<tr>
<td>Stockpiling</td>
<td>171,421 (3)</td>
<td></td>
</tr>
<tr>
<td>Relaxation Techniques</td>
<td>164,262 (3)</td>
<td></td>
</tr>
<tr>
<td>Excess Sleep</td>
<td>127,623 (2)</td>
<td></td>
</tr>
<tr>
<td>Pets</td>
<td>109,510 (2)</td>
<td></td>
</tr>
<tr>
<td>Postponing Plans</td>
<td>98,626 (2)</td>
<td></td>
</tr>
<tr>
<td>Childcare</td>
<td>97,735 (2)</td>
<td></td>
</tr>
<tr>
<td>Public Transportation</td>
<td>94,414 (2)</td>
<td></td>
</tr>
<tr>
<td>Reduced Sleep Quality</td>
<td>88,196 (1)</td>
<td></td>
</tr>
<tr>
<td>Home School</td>
<td>80,153 (1)</td>
<td></td>
</tr>
<tr>
<td>Non-COVID Hospital Visits</td>
<td>77,278 (1)</td>
<td></td>
</tr>
<tr>
<td>Doctor Well Visit</td>
<td>72,235 (1)</td>
<td></td>
</tr>
<tr>
<td>Funerals</td>
<td>45,394 (1)</td>
<td></td>
</tr>
<tr>
<td>Family-Centered Time</td>
<td>28,106 (0)</td>
<td></td>
</tr>
<tr>
<td>Outdoor Culture</td>
<td>21,546 (0)</td>
<td></td>
</tr>
<tr>
<td>Births</td>
<td>17,283 (0)</td>
<td></td>
</tr>
<tr>
<td>Telehealth</td>
<td>11,574 (0)</td>
<td></td>
</tr>
<tr>
<td>Smokeless Tobacco Consumption</td>
<td>2,136 (0)</td>
<td></td>
</tr>
<tr>
<td><strong>COVID-19 Mentions (37)</strong></td>
<td><strong>3,390,139</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Number of posts by source type (June 15 to November 15)
Within the data sample, 6,210,255 (69%) posts were classified as belonging to the category of daily life topics, 3,390,139 (37%) contained mentions of COVID-19, and 1,836,200 (20%) posts were classified as belonging to the category of COVID-19-related public health topics (Table 1). The most prevalent topics among the daily life posts were Sex Life (887,457 [14%]), Food (838,513 [14%]), and Financial Concerns (710,757 [11%]). The most prevalent topic in COVID-19-related public health behaviors posts was Wearing Face Masks (1,120,344 [61%]), followed by Lockdown (457,705 [25%]), and Social Distancing (242,105 [13%]).

Online social posts including COVID-19 mentions and COVID-19-related public health behaviors increased in June, as COVID-19 cases also increased, but remained stagnant as cases began to increase in October (Figure 1). Discussion about wearing face masks was most prevalent in mid-July, during the summer wave (mid-June to early September) of COVID-19 cases and has remained at pre-June levels in October and November, with the exception of a sharp increase on October 2, 2020 (Figure 2).

**Discussion**

From June to November 2020, predominantly US-based online social chatter was more focused on daily life than it was on public health behaviors relating to COVID-19. Although discussion relating to COVID-19 and related public health behaviors appeared to increase with rising US cases in the summer wave (early June to early September), the volume of COVID-19-related discussion was lower in the ongoing wave that began in the fall (mid-October), despite the fact that during the fall wave, COVID-19 cases increased to their highest rates since the
pandemic began.[36] In particular, discussion of wearing facemasks, the most prevalent of any COVID-19 public health behavior we studied, declined in mid-July despite the pandemic continuing and evidence that wearing facemasks has not been universally adopted in the US, and increased only minimally once cases began to increase again in early October.[37,38] One exception was the brief but stark increase in COVID-19-related discussion on October 2, 2020, which coincided with the announcement that President Donald Trump had contracted COVID-19.[39] The high prevalence of daily life topics in social media chatter compared with COVID-19-related public health behaviors and mentions of COVID-19 is not immediately surprising given the differences in scope; however, the decrease of COVID-19 related discussion in the context of rising COVID-19 cases differs from the pattern we visualized in the summer wave and from patterns reported during the spring (March to June) wave.

Our findings differed from those of previous COVID-19-related social media analyses using Twitter and conducted earlier in the pandemic. Reports from March that used data from Twitter indicated that social media discussion about COVID-19-related health discussion was high, and more common than discussion about daily life topics such as socializing, the economy, or politics.[40] Earlier research also found that COVID-19-related public health measures were discussed more often than social topics and more often than other COVID-19-related topics.[15,7] Although this difference may be due to different sources and time periods, a change might have occurred in the public’s focus on COVID-19 preventative behaviors. For instance, as public health experts have warned against relaxing preventive behaviors as pandemic fatigue builds, activity and traffic data have indicated that people may have stopped adhering to public health recommendations to stay home and avoid close contact.[41-44] The decline of chatter regarding wearing facemasks, and the relative low rates of discussions on other COVID-19-
related public health behaviors may reflect that social media engagement with these issues has
decreased as the pandemic has progressed, and remains low among the US population as the
pandemic continues to confront a high COVID-19 daily case rate.

Our study has several limitations. First, although our third-party data provider reported
that about 70% of posts were from the US, we do not know the location for most posts according
to our direct geotagging methods, which were only able to tag about 80% of posts (eTable 2). As
a result, we cannot make international comparisons, but our dataset is more representative of the
US than of any other country. Second, the number of posts included in our dataset was much
lower than previous studies, likely due to the types of data sources used, which excluded sites
such as Twitter in order to exclude noise that might have obscured signals in social media data,
and our methodology, which included removing posts not relevant to our more refined
taxonomy. We used a stringent exclusion criterion with a list of prespecified keywords that may
also have led to a smaller sample size, but our approach aimed to create a sample with high
specificity. Finally, there is no demographic information available from the data posts directly
due to privacy considerations and data use agreements. Thus, we cannot determine whether our
data sample contains biases due to the demographics of the people who post. For instance,
Reddit, which was the most common forum source for our data sample, has been found to be
used by a younger, male audience.[45,46]

Conclusion

In this study of predominantly US-based COVID-19 social media data from June to
November 2020, we observed that COVID-19 and relevant public health measures were
discussed less than daily life behaviors on social media, and that discussion on wearing
facemasks decreased throughout the summer and into the fall, while cases increased. These
discussion rates may reveal a need for increased public health messaging as the pandemic continues.

Acknowledgments

Alina Cohen, Tali Moed, Pini Matzner, and Yahel Oren from Signals Analytics had full access to the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Daisy Massey from Yale School of Medicine takes full responsibility for the data interpretation and writing. This work was supported by the project *Insights about the COVID Pandemic Using Public Data* IRES PD: 20-005872 with funding from the Foundation for a Smoke-Free World.

Conflicts of Interest

Yuan Lu is supported by the National Heart, Lung, and Blood Institute (K12HL138037) and the Yale Center for Implementation Science. Rachel Dreyer is supported by an American Heart Association Transformational Project Award (#19TPA34830013) and a Canadian Institutes of Health Research Project Grant (RN356054–401229). In the past three years, Harlan Krumholz received expenses and/or personal fees from UnitedHealth, IBM Watson Health, Element Science, Aetna, Facebook, the Siegfried and Jensen Law Firm, Arnold and Porter Law Firm, Martin/Baughman Law Firm, F-Prime, and the National Center for Cardiovascular Diseases in Beijing. He is an owner of Refactor Health and HugoHealth, and had grants and/or contracts from the Centers for Medicare & Medicaid Services, Medtronic, the U.S. Food and Drug Administration, Johnson & Johnson, and the Shenzhen Center for Health Information. The remaining authors have no disclosures to report.
Abbreviations

COVID-19: coronavirus disease 2019

US: United States

API: application programming interface

NLP: natural language processing

EVALI: e-cigarette or vaping use-associated lung injury
References


46. Finlay SC. Age and gender in Reddit commenting and success. 2014. doi: 10.1633/JISTaP.2014.2.3.2
Figure 1 Online Social Discussion Categories vs. US Daily New COVID-19 Cases (June 15 to November 15, 2020)

Figure 2 Public Health Measures Online Social Discussion vs. US Daily New COVID-19 Cases (June 15 to November 15, 2020)