JMIR Publications
The Leading eHealth Publisher

Tracking self-reported symptoms and medical conditions on social media during the COVID-19 pandemic

Qinglan Ding, Daisy Massey, Chenxi Huang, Connor Grady, Yuan Lu, Alina Cohen, Pini Matzner, Shiwani Mahajan, César Caraballo, Navin Kumar, Yuchen Xue, Rachel Dreyer, Brita Roy, Harlan M. Krumholz

Submitted to: JMIR Public Health and Surveillance
on: April 06, 2021

Table of Contents

Tracking self-reported symptoms and medical conditions on social media during the COVID-19 pandemic

Qinglan Ding[1] MBBS, PhD; Daisy Massey[2] BA; Chenxi Huang[2] PhD; Connor Grady[3] BSc; Yuan Lu[2, 4] SCD; Alina Cohen[5] MSc, MBA; Pini Matzner[5] PhD; Shiwani Mahajan[2, 4] MBBS, MHS; César Caraballo[2, 4] MD; Navin Kumar[6, 7] MPhil; Yuchen Xue[8] MA; Rachel Dreyer[9] PhD; Brita Roy[3, 10] MD, MPH, MHS; Harlan M. Krumholz[4, 1, 11] MD, MS

[1]College of Health and Human Sciences Purdue University West Lafayette US
[2]Center for Outcomes Research and Evaluation Yale New Haven Hospital New Haven US
[3]Department of Chronic Disease Epidemiology Yale School of Public Health New Haven US
[4]Section of Cardiovascular Medicine Department of Internal Medicine Yale School of Medicine New Haven US
[5]Signals Analytics Netanya IL
[6]Department of Sociology Yale University New Haven US
[7]Institute for Network Science Yale University New Haven US
[8]Foundation for a Smoke-Free World New York US
[9]Department of Emergency Medicine Yale School of Medicine New Haven US
[10]Department of Medicine Yale School of Medicine New Haven US
[11]Department of Health Policy and Management Yale School of Public Health New Haven US

Corresponding Author:
Harlan M. Krumholz MD, MS
College of Health and Human Sciences
Purdue University
West Lafayette
US

Abstract

Background: Harnessing health-related data posted on social media in real-time has the potential to offer insights into how the pandemic impacts the mental health and general well-being of individuals and populations over time.

Objective: The aim of this study was to obtain information on symptoms and medical conditions self-reported by non-Twitter social media users during the coronavirus disease 2019 (COVID-19) pandemic, and to determine how discussion of these symptoms and medical conditions on social media changed over time.

Methods: We used natural language processing (NLP) algorithms to identify symptom and medical condition topics being discussed on social media between June 14 and December 13, 2020. The sample social media posts were geotagged by NetBase, a third-party data provider. We calculated the positive predictive value and sensitivity to validate the classification of the posts. We also assessed the frequency of different health-related discussions on social media over time during the study period, and compared the changes in the frequency of each symptom/medical condition discussion to the fluctuation of U.S. daily new COVID-19 cases during the study period. Additionally, we compared the trends of the 5 most commonly mentioned symptoms and medical conditions from June 14 to August 31 (when the U.S. passed 6 million COVID-19 cases) to the trends observed from September 1 to December 13, 2020.

Results: Within a total of 9,807,813 posts (nearly 70% were sourced from the U.S.), we identified discussion of 120 symptom topics and 1,542 medical condition topics. Our classification of the health-related posts had a positive predictive value of over 80% and an average classification rate of 92% sensitivity. The 5 most commonly mentioned symptoms on social media during the study period were: anxiety (in 201,303 posts or 12.2% of the total posts mentioning symptoms), generalized pain (189,673, 11.5%), weight loss (95,793, 5.8%), fatigue (91,252, 5.5%), and coughing (86,235, 5.2%). The 5 most discussed medical conditions were: COVID-19 (in 5,420,276 posts or 66.4% of the total posts mentioning medical conditions), unspecified infectious disease (469,356, 5.8%), influenza (270,166, 3.3%), unspecified disorders of the central nervous system (253,407, 3.1%), and depression (151,752, 1.9%). The changes in the frequency of 2 medical conditions, COVID-19 and unspecified infectious disease, were similar to the fluctuation of daily new confirmed cases of COVID-19 in the U.S.

Conclusions: COVID-19 and symptoms of anxiety were the two most commonly discussed health-related topics on social media from June 14 to December 13, 2020. Real-time monitoring of social media posts on symptoms and medical conditions may help assess the population's mental health status and enhance public health surveillance for infectious disease.

(JMIR Preprints 06/04/2021:29413)
DOI: https://doi.org/10.2196/preprints.29413

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Tracking self-reported symptoms and medical conditions on social media during the COVID-19 pandemic

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

[1] Center for Outcomes Research and Evaluation, Yale New Haven Hospital, New Haven, Connecticut; [2] College of Health and Human Sciences, Purdue University, West Lafayette, Indiana; [3] Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, Connecticut; [4] Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut; [5] Signals Analytics, New York, New York and Netanya, Israel; [6] Department of Sociology, Yale University, New Haven, Connecticut;

[7] Foundation for a Smoke-Free World, New York, New York; [8] Department of Emergency Medicine, Yale School of Medicine, New Haven, Connecticut; [9] Department of Medicine, Yale School of Medicine, New Haven, Connecticut; [10] Department of Health Policy and Management, Yale School of Public Health, New Haven, Connecticut

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 (n=419 words, Max 450 words)

Background: Harnessing health-related data posted on social media in real-time has the potential to offer insights into how the pandemic impacts the mental health and general well-being of individuals and populations over time.

Objective: The aim of this study was to obtain information on symptoms and medical conditions self-reported by non-Twitter social media users during the coronavirus disease 2019 (COVID-19) pandemic, and to determine how discussion of these symptoms and medical conditions on social media changed over time.

Methods: We used natural language processing (NLP) algorithms to identify symptom and medical condition topics being discussed on social media between June 14 and December 13, 2020. The sample social media posts were geotagged by NetBase, a third-party data provider. We calculated the positive predictive value and sensitivity to validate the classification of the posts. We also assessed the frequency of different health-related discussions on social media over time during the study period, and compared the changes in the frequency of each symptom/medical condition discussion to the fluctuation of U.S. daily new COVID-19 cases during the study period. Additionally, we compared the trends of the 5 most commonly mentioned symptoms and medical conditions from June 14 to August 31 (when the U.S. passed 6 million COVID-19 cases) to the trends observed from September 1 to December 13, 2020.

Results: Within a total of 9,807,813 posts (nearly 70% were sourced from the U.S.), we identified discussion of 120 symptom topics and 1,542 medical condition topics. Our classification of the health-related posts had a positive predictive value of over 80% and an average classification rate of 92% sensitivity. The 5 most commonly mentioned symptoms on social media during the study period were: anxiety (in 201,303 posts or 12.2% of the total posts mentioning symptoms), generalized pain (189,673, 11.5%), weight loss (95,793, 5.8%), fatigue (91,252, 5.5%), and coughing (86,235, 5.2%). The 5 most discussed medical conditions were: COVID-19 (in 5,420,276 posts or 66.4% of the total posts mentioning medical conditions), unspecified infectious disease (469,356, 5.8%), influenza (270,166, 3.3%), unspecified disorders of the central nervous system (253,407, 3.1%), and depression (151,752, 1.9%). The changes in the frequency of 2 medical conditions, COVID-19 and unspecified infectious disease, were similar to the fluctuation of daily new confirmed cases of COVID-19 in the U.S.

Conclusions: COVID-19 and symptoms of anxiety were the two most commonly discussed health-related topics on social media from June 14 to December 13, 2020. Real-time monitoring of social media posts on symptoms and medical conditions may help assess the population's mental health status and enhance public health surveillance for infectious disease.

Keywords: Health conditions; symptoms; mental health; social media; infoveillance; public health surveillance; COVID-19; pandemic; natural language processing

Introduction

The coronavirus disease 2019 (COVID-19) is continuing its spread across the globe, with more than 131 million confirmed cases and 2,862,885 deaths in 188 countries as of April 6, 2021.[1] As individuals are being encouraged to telecommute and self-quarantine, social media usage has surged by over 40%, emerging as a powerful tool for facilitating communication and disseminating information in a timely manner.[2, 3] The general public and health care professionals use social media platforms for health surveillance, to share their feelings, opinions, knowledge, and experiences in relation to the COVID-19 pandemic, and interact with others who share similar characteristics or interests.[4-7] A growing number of people also use social media to seek and share health information that might otherwise be “invisible” to clinicians and medical researchers (e.g., self-diagnosis and self-treated symptoms with over-the-counter medications).[8-10] Harnessing publicly-available health-related data posted on social media in real time has the potential to offer insights into how the pandemic impacts the mental health and general well-being of individuals and populations over time.[2, 11]

Although prior studies have demonstrated that social media discussions can influence health-related beliefs and behaviors, more studies are needed to understand how social media plays a role during the pandemic.[12, 13] Since the emergence of the COVID-19 pandemic, an estimated 41% of U.S. adults have delayed or avoided urgent and routine medical care during the pandemic due to concerns about COVID-19.[14] Real-time information regarding self-reported health status at a population level is lacking. Most literature in this area of research has been focused particularly on mental health or COVID-19 symptoms, with Twitter frequently being utilized as the sole data source.[15, 16] There was limited information regarding health-related discussions from social media sites other than Twitter. Furthermore, the COVID-19 infection predictive value of social media discussions has not yet been ascertained.[17] Extracting and analyzing health-related data from multiple social media sources might provide novel ways of measuring the health status and the full spectrum of symptoms and illness of the population in real-time.[11, 18]

As such, we created a dashboard to extract and monitor posts mentioning symptoms and medical conditions from social media sites other than Twitter over the course of the COVID-19 pandemic. In this study, we sought to answer the following questions: (1) What symptoms and medical conditions were people talking about on social media other than Twitter during the COVID-19 pandemic? (2) How have discussions of symptoms and medical conditions on social media changed over a 6-month period during the pandemic? (3) Were daily fluctuations in health-related social media conversations associated with daily changes in new confirmed cases of COVID-19 in the U.S.?

Methods

Data Collection

We included English-language social networks and forums worldwide, such as Facebook public pages, Reddit, 4Chan and comment sections of news sites, such as ABC news.[19] We defined forums as thread-based message boards and topic-specific pages.[20] Twitter was excluded because of the study purpose and our concerns about the high bot volume and decreased context related to Twitter’s character limits and hashtag-based posts.[21, 22] In this study, we preferred sources that had longer posts with more context so that natural language processing (NLP) classification was more accurate. We partnered with Signals Analytics, an advanced analytics company, to obtain access to target data sources from a third-party data vendor (NetBase), and to conduct the analytics.[23, 24] In order to geotag posts, NetBase used a combination of geotagged social media messages, author profiles, and each country’s unique website domain suffix (e.g., .ca for Canada). All the acquired data were then deidentified by NetBase and transferred to Signals Analytics for analysis.

We also gathered data for COVID-19 cases from the COVID-19 Dashboard developed by the Center for System Science and Engineering at Johns Hopkins University, which provides the most comprehensive and up-to-date information on COVID-19 trends.[1] Using the RapidAPI application programming interface (API),[25] we updated the COVID-19 statistics (daily new cases, or incidence) on a daily basis.

In this study, all personal identifying information such as usernames, emails, and IP addresses were removed before analysis. The study was exempt from Institutional Review Board review at Yale University as it used publicly available, anonymized data.

Data Analysis

For the data analysis of symptoms and medical conditions being discussed on social media between June 14 (when many countries began to lift major COVID-19 restrictions) and December 13, 2020 (when the first shipment of the COVID-19 vaccine arrived in the U.S.), we began by applying NLP algorithms to process social media posts collected from data sources during the study period, and then classified these posts according to symptoms and medical conditions being mentioned.

To do this, NetBase ran a daily scheduled data extraction query that we designed for the study on over 300 million online data sources (Supplement 1). Additionally, we performed the following filtering steps to include posts relevant to our research questions. First, NLP algorithms were run, and advertisements and posts on sites for pornography were removed

(Supplement 2). Next, we applied a taxonomy of over 3,000 health-related topics to identify key words, phrases, and statements mentioning symptoms and medical conditions (Supplement 3). Social media posts that did not contain any of the taxonomy terms or symptoms and medical conditions keywords were then deleted. Lastly, we removed redundant posts, blog posts and news articles to ensure that the analysis was based on unique posts from social networks, forums, and comments only.

To evaluate the performance of the NLP algorithms and taxonomy classifications of symptoms and medical conditions, we applied the taxonomy to an independent sample of 100 posts and calculated the positive predictive value and sensitivity of the classification (Supplement 4). Additionally, we validated our methodology by applying the algorithm to detect major news events that occurred during the study period. We were able to observe a dramatic increase in the volume of online discussion on topics relating to the event immediately following the occurrence (Supplement Figure 1a &1b). The methodology used in our study has also been previously used to provide insights into the characterization and prediction of e-cigarette or vaping product use-associated lung injury outbreak known as the EVALI study.[26]

Our taxonomy was organized into three levels: categories, subcategories and topics. Symptoms and medical conditions were the two main categories in the taxonomy. (Table 1 and Supplement 3). The symptoms category included 98 non-COVID-19 topics (symptoms), which were grouped into 7 subcategories based on the affected organ or systems (e.g., cardiovascular, respiratory, etc.) A list of 22 COVID-19 related topics (symptoms) was included as a separate symptom subcategory. The list of COVID-19 related symptoms was defined as outlined by the Center for Disease Control and Prevention (CDC) on December 22, 2020. [27] Because our algorithms captured all posts that mentioned any of the listed COVID-19 symptoms in the

COVID-19 related symptom subcategory, the included posts may not necessarily represent discussions of symptoms experienced by COVID-19 patients. The medical condition category included 2,200 topics (medical diagnoses), which were grouped into 10 subcategories. Categories, subcategories and topics in the taxonomy were not mutually exclusive; each post could be assigned to multiple categories, subcategories or topics.

We also created content filters to retain posts mentioning COVID-19 for further analysis. We applied two filters: COVID-19 disease status and COVID-19 diagnostic methods to identify discussions of COVID-19 disease status (tested positive or negative, symptomatic or asymptomatic, recovered, and exposed to a confirmed patient), and diagnostic methods (COVID-19 testing, self-diagnosed, and remotely diagnosed). These more restrictive searches were conducted by activating the two additional filters using the NLP algorithm, and the resulting posts from that search may not indicate the author's COVID-19 status.

To explore how discussion of symptoms and medical conditions on social media changed from June 14 to December 13, 2020, we determined the number of posts that included discussion of each symptom and medical condition over time using NLP classification (Supplement 5). In order to assess whether changes in frequency of health-related social media conversations were associated with daily fluctuation in COVID-19 cases, we also visualized the U.S. daily new COVID-19 cases and changes in discussions of the top 5 most-talked-about symptom and medical condition topics using time-series plots.

Additionally, we compared the trends of the 5 most mentioned symptoms and medical conditions from June 14 to August 31 (when the U.S. passed 6 million COVID-19 cases) to the trends observed from September 1 to December 13, 2020, by measuring the percent change between the two time periods in the number of posts including discussion of each topic. We compared the two time periods in order to reveal changes in health-related conversations on social media at different stages of the pandemic, as prior literature focused primarily on the early stage of the pandemic (before June 2020). Our approach was also designed to contribute to a better understanding of the impact of COVID-19 on the public’s perceptions and attitudes toward different symptoms, medical conditions, and health care seeking behaviors.

Results

After social media posts were collected from sources, pre-processed, and classified according to the taxonomy by NLP algorithms, our final sample included a total of 9,807,813 posts between June 14 and December 13, 2020, that mentioned at least one of the 120 symptoms or 1,542 medical condition topics in our taxonomy (Table 1). Our taxonomy classification in the independent sample of 100 posts resulted in a positive predictive value of over 80% and an average classification rate of 92% sensitivity. Also, according to indirect geotagging information provided by NetBase, about 70% of all posts collected by the search query were from the U.S. The most prevalent symptom subcategory was "neuro-psychological symptoms" (34.5%), followed by COVID-19 related symptoms subcategory (30.4%). The most prevalent medical condition subcategory was "infectious disease" (74.2%), followed by the subcategory of "psychiatric or mental health disorders" (6.0 %) (Table 1).

Table 1. Number of symptom and medical condition posts mentioned on social media by taxonomy topic (June 14 to December 13, 2020)

Relevant Taxonomy Categories and Subcategories (number of topics), N total=9,807,813	Number of Posts with Symptoms or Medical Conditions	Percentage out of All Symptoms or All Medical Conditions Posts (%)
Symptoms (n=1,649,547)		
Neuropsychological Symptoms (17)	568,662	34.47%
COVID-19 related Symptoms*	501,178	30.38%

(22)		
Respiratory Symptoms (7)	128,134	7.77%
Gastrointestinal Symptoms (13)	120,621	7.31%
Dermal Symptoms (16)	99,453	6.03%
CVD Symptoms (4)	34,014	2.06%
Musculoskeletal Symptoms (7)	33,604	2.04%
Other Symptoms (34)	163,881	9.93%
Medical Conditions (n=8,158,266)		
Infectious Disease (80)	6,052,068	74.18%
Psychiatric or Mental Health Disorders (21)	484,505	5.94%
Neurovascular & Cardiovascular Diseases (63)	465,675	5.71%
Respiratory Disorders (17)	165,404	2.03%
Hematological & Oncological Disorders (127)	164,159	2.01%
Other Disorders (1234)	828,786	10.13%

* COVID-19 related symptoms were based on CDC symptoms of COVID-19 updated on December 22, 2020. The included 22 COVID-19 symptoms were: runny nose, change in sense of taste, change in sense of smell, chills, bluish lips/face, inability to stay awake, fatigue, headache, sore throat, abdominal pain, vomiting, muscle pain/spasms, drowsiness, nausea, body aches, chest pain, itching/swelling, fever, confusional state, diarrhea, coughing, and difficulty breathing.

Irrespective of subcategories classification, the five most commonly mentioned symptom topics were anxiety (201,303, 12.20% out of total posts mentioning symptoms), generalized pain (189,673, 11.5%), weight loss (95,793, 5.8%), fatigue (91,252, 5.5%), and coughing (86,235, 5.2%), accounting for 40.2 % of all symptom posts combined (Table 2 and Supplement Figure 2). The 5 most discussed medical condition topics were COVID-19 (5,420,276, 66.4% of the total posts mentioning medical conditions), unspecified infectious disease (469,356, 5.8%), influenza (270,166, 3.3%), unspecified disorders of the central nervous system (CNS) (253,407, 3.1%), and depression (151,752, 1.9%), and they combined accounted for 80.5% of all medical conditions discussed on social media during the study period (Table 2 and Supplement Figure 3).

Table 2. Frequency of top 5 most discussed symptoms and medical conditions on social media by taxonomy topic (June 14 to December 13, 2020).

Relevant Taxonomy Categories and Topics, N total=9,807,813	Number of Posts with Symptoms or Medical Conditions Topics	Percentage out of All Symptoms or All Medical Conditions Topics Posts (%)
Symptoms (n=1,649,547)		
Anxiety	201,303	12.20%
Generalized pain	189,673	11.49%
Weight loss	95,793	5.81%
Fatigue	91,252	5.53%
Coughing	86,235	5.23%
Medical Conditions (n=8,158,266)		
COVID-19[a]	5,420,276	66.44%
Unspecified infectious disease	469,356	5.75%
Influenza	270,166	3.31%
Unspecified CNS[b] disorders	253,407	3.11%
Depression	151,752	1.86%

Abbreviations: a, coronavirus disease 2019; b, CNS-central nervous system

Within the COVID-19 related symptoms subcategory, fatigue (91,208, 32.9%) and coughing (86,222, 31.1%) were the most-talked-about COVID-19 related symptom topics (Table 3). Bluish lips/face (1,019, 0.4%) and inability to stay awake (486, 0.2%) were the least commonly discussed COVID-19 symptoms.

Table 3. Comparing changes in number of COVID-19 symptom posts between June 14 and August 31, 2020 with posts number during September 1-December 13, 2020.

COVID-19 Related Symptoms Per CDC definition*	**Total number of posts mentioning this (% of all COVID symptoms posts)**	**Number of posts during June 14-August 31, 2020**	**Number of posts during September 1-December 13, 2020**	**% changes in number of posts**
Fatigue	91,208 (32.88)	36,876	54,332	47.33
Coughing	86,222 (31.08)	41,163	45,059	9.46
Fever	59,906 (21.59)	27,729	32,177	16.04
Headache	41,693 (15.02)	18,052	23,641	30.96
Vomiting	39,103 (14.09)	17,364	21,739	25.19
Difficulty Breathing	33,589 (12.11)	16,917	16,672	Decreased 1.45
Nausea	29,103 (10.49)	13,039	16,064	23.19

Itching/Swelling	28,337 (10.22)	12,953	15,384	18.77
Sore Throat	14,694 (5.29)	6,424	8,270	28.74
Diarrhea	14,140 (5.09)	6,716	7,424	10.54
Chest pain	9,412 (3.39)	4,255	5,157	21.19
Abdominal Pain	9,238 (3.33)	4,080	5,158	26.42
Runny Nose	8,283 (2.98)	3,029	5,254	73.46
Body Aches	7,871 (2.84)	3,540	4,331	22.34
Change in Sense of Taste	6,510 (2.35)	2,447	4,063	66.04
Muscle Pain/Spasms	6,321 (2.28)	2,816	3,505	24.47
Change in Sense of Smell	6,192 (2.23)	2,340	3,852	64.62
Confusional State	3,716 (1.34)	1,737	1,979	13.93
Chills	2,879 (1.04)	1,141	1,738	52.32
Drowsiness	1,256 (0.45)	560	696	24.29
Bluish Lips/Face	1,019 (0.37)	404	615	52.23
Inability to stay awake	486 (0.18)	195	291	49.23

*The COVID-19 symptoms list was updated December 22, 2020 from the CDC update. Please note: Our algorithms captured all posts mentioning any of these symptoms in the COVID-19 symptom subcategory. As a result, the posts may not necessarily represent patients discussing their own COVID-19 symptoms.

After applying the COVID-19 disease status filter to all posts mentioning the top 5 most mentioned symptoms and medical conditions, we noticed that within the posts classified with the medical condition of COVID-19, 62.9% had also discussed tested positive, and 9.1% of the discussions were related to asymptomatic COVID-19 (Table 4). Applying the COVID-19 diagnostic method filter revealed that the most popular COVID-19 diagnostic methods discussed were COVID-19 tests regardless of symptom or medical condition subcategory. (Table 4).

Table 4. Frequency of COVID-19 disease status and diagnostic methods discussions for the 5 most commonly mentioned symptoms and medical conditions on social media (June 14 to December 13, 2020, not indicative of a person's COVID-19 status)

COVID-19 Filters	**Five Most Commonly Mentioned Symptoms**				
	Anxiety	Generalized Pain	Weight Loss	Fatigue	Cough
COVID-19 disease status (Percentage of posts out of total # of posts mentioning this for a specific symptom)					
Tested positive	2,202 (37.3%)	1,669 (37.8%)	220 (43.7%)	3,402 (41.4%)	7,916 (37.3%)
Tested negative	779 (13.2%)	771 (17.5%)	59 (11.7%)	1,412 (17.2%)	3,034 (14.3%)
Symptomatic	1,531 (25.9%)	742 (16.8%)	82 (16.3%)	1,585 (19.3%)	4,092 (19.3%)
Asymptomatic	869 (14.7%)	850 (19.3%)	79 (15.7%)	1,257 (15.3%)	4,920 (23.2%)
Exposed to confirmed patient	407 (6.9%)	227 (5.1%)	25 (4.9%)	350 (4.3%)	1,038 (4.9%)
Recovered	123 (2.1%)	156 (3.5%)	38 (7.6%)	206 (2.5%)	205 (1.0%)
COVID-19 diagnostic methods (Percentage of posts out of total # of posts mentioning this for a specific symptom)					
COVID-19 tests	2,981 (97.6%)	2,440 (96.5%)	279 (97.2%)	4,814 (99.0%)	10,950 (99.3%)

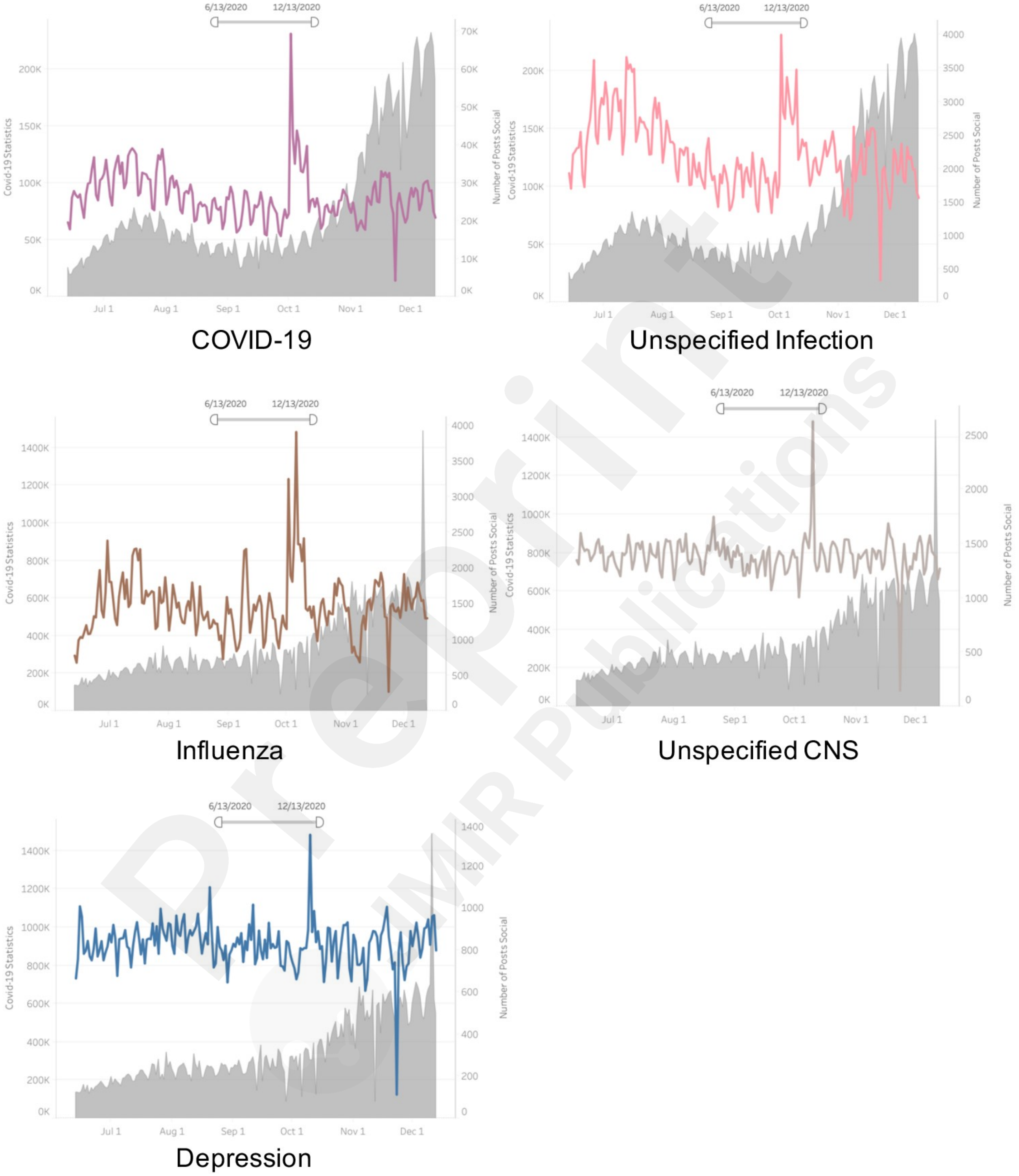
Remotely diagnosed	31 (1.0%)	45 (1.8%)	7 (2.5%)	27 (0.6 %)	51 (0.5%)
Self-diagnosed	43 (1.4%)	43 (1.7%)	1 (0.3%)	21 (0.4%)	31 (0.3%)
COVID-19 Filters	**Five Most Commonly Mentioned Medical Disorders**				
	COVID-19	Unspecified Infection	Influenza	Unspecified CNS	Depression
COVID-19 disease status (percentage of posts out of total # of posts mentioning this for a medical disorder)					
Tested positive	503,564 (62.9%)	33,543 (46.5%)	12,349 (40.4%)	574 (38.0%)	662 (40.8%)
Tested negative	94,063 (11.8%)	7,672 (10.6%)	4,565 (14.9%)	141 (9.3%)	193 (11.9%)
Symptomatic	64,065 (8.0%)	10,073 (13.9%)	4,657 (15.2%)	325 (21.5%)	331 (20.4%)
Asymptomatic	72,985 (9.1%)	16,008 (22.2%)	6,944 (22.7%)	299 (19.8%)	214 (13.1%)
Exposed to confirmed patient	60,208 (7.5%)	4,413 (6.1%)	1,883 (6.2%)	94 (6.3%)	134 (8.3%)
Recovered	5,123 (0.6%)	366 (0.5%)	158 (0.5%)	77 (5.1%)	87 (5.4%)
COVID-19 diagnostic methods (percentage of posts out of total # of posts mentioning this for a specific medical disorder)					
COVID-19 tests	597,627 (98.9%)	41,215 (99.4%)	16,914 (99.5%)	715 (97.5%)	855 (96.8%)
Remotely diagnosed	6,408 (1.1%)	202 (0.5%)	48 (0.3%)	13 (1.8%)	9 (1.0%)
Self-diagnosed	530 (0.1%)	41 (0.1%)	37 (0.2%)	5 (0.7%)	19 (2.2%)

When examining changes in the frequency of the top 5 most commonly mentioned symptom topic discussions over the 6-month study period, we noted a 24% increase in symptom posts mentioning anxiety, generalized pain, and fatigue during September 1-December 13, 2020 (vs. June 14-August 31, 2020) (Supplement Figure 4). Compared with June 14-August 31, 2020, posts mentioning the medical condition topics influenza, unspecified CNS disorders, and depression increased by more than 27% during September 1-December 13, 2020 (Supplement

Figure 5). In terms of changes within the COVID-19 related symptoms subcategory, social media posts mentioning runny nose, change in the sense of taste and smell increased over 64%, while posts mentioning difficulty breathing decreased 1.5% during September 1-December 13, 2020 (vs. June 14-August 31, 2020) (Supplement Figure 6).

In analyzing the associations between changes in posts mentioning symptoms or medical conditions and daily fluctuations of COVID-19 statistics, we found that patterns of changes in posts frequency mentioning both COVID-19 and unspecified infectious disease appeared to be similar to the trend of daily new confirmed cases in the U.S. over time, especially from June 14 to October 2, 2020 (Figure 1). The pattern of changes in posts frequency mentioning other medical conditions or symptoms did not appear similar to fluctuation of daily COVID-19 statistics (Figures 1 and 2). Additionally, we noticed a significant increase in daily frequency of posts mentioning the top 5 symptom and medical condition topics in October and a decrease in late November-December 2020 (Supplement Table 1).

Figure 1. Associations between changes in U.S. new daily COVID-19 cases and the number of medical conditions posts (June 13-December 13, 2020).



COVID-19

Unspecified Infection

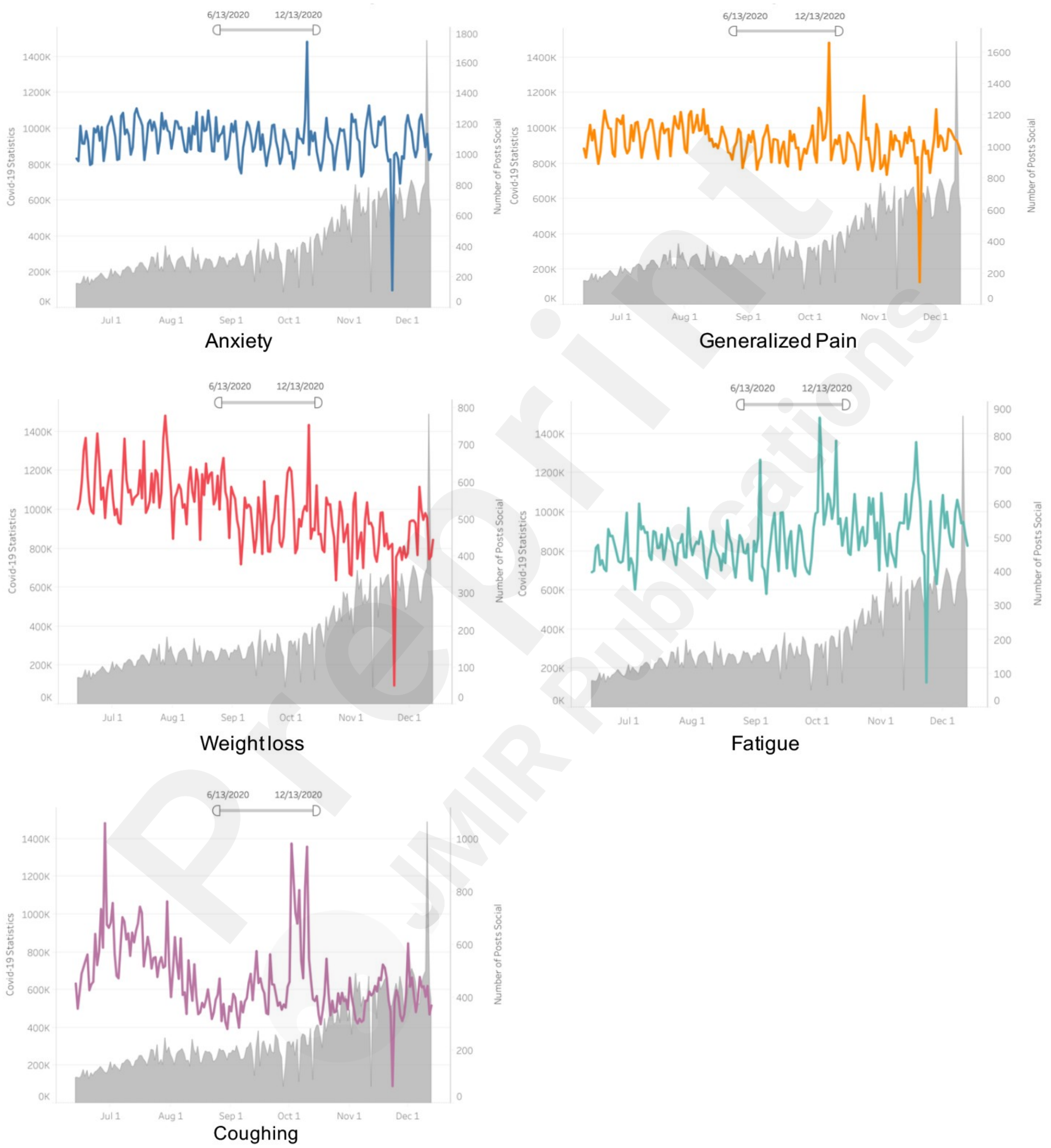
Influenza

Unspecified CNS

Depression

Note: the gray shaded area indicated daily active COVID-19 cases in the U.S., while the colored curves showed fluctuations in posts mentioning different medical disorders during the study period.

Figure 2. Associations between changes in U.S. new daily COVID-19 cases and the number of symptoms posts (June 13-December 13, 2020).



Anxiety

Generalized Pain

Weight loss

Fatigue

Coughing

Note: the gray shaded area indicated daily active COVID-19 cases in the U.S., while the colored curves showed fluctuations in posts mentioning different symptoms during the study period.

Discussion

In this study, we collected and analyzed online posts from forums and comments on news sites between June 14 and December 13, 2020. We found that a wide variety of symptoms and medical conditions topics were discussed on social media. While the vast majority of discussions were about COVID-19 infection and COVID-19 related symptoms (as defined by CDC), neuropsychological symptoms (e.g., anxiety) and other medical conditions (e.g., infectious disease and psychiatric disorders) were also frequently mentioned. As COVID-19 cases continued to rise globally, the cumulative volume of posts mentioning anxiety, generalized pain, fatigue, influenza, unspecified CNS disorders, and depression increased from September 1 to December 13 (compared with June 13 to August 31, 2020). Additionally, we noticed that the pattern of changes in the frequency of daily online conversations about COVID-19 infection and unspecified infectious disease appeared to be similar to the trend of daily new confirmed cases of COVID-19 in the U.S. over time.

Our findings expand on previous observations regarding the mental health effects of the COVID-19 pandemic among social media users by presenting a more complete picture of health-related topics discussed on social media.[11, 28] Our results not only confirm the findings from previous studies that showed high levels of anxiety and depression mentioned by social media users during the pandemic,[29, 30] but also indicate that a broader range of symptoms and medical conditions, including rare diseases, was mentioned on social media. These data support the idea that social media represents a possible powerful source of information for health care professionals to draw real-time estimations about population health status.[18, 28] As access to the internet becomes more widely available and with the anonymity of social media, people who face barriers to accessing health care and those who have mental health symptoms may use social media to speak openly about their health experiences and seek help.[18, 31] Collectively, these results further justify our approach to monitoring symptoms and medical conditions posts on social media during the pandemic, and call for further investigation of the possibility of using social media analytics to gain insights into the population's symptoms, including mental health symptoms, which are difficult to monitor outside of the health system, health threats and to enhance public health preparedness.

As the pandemic progresses, obtaining information on the symptoms profile of COVID-19 could help to better diagnose and treat the disease. There has been increasing recognition of the importance of extracting social media information to explore symptom experience and disease progression among COVID-19 patients.[32] Although we did not restrict our analysis to only social media posts containing COVID-19 and could not verify the authors' disease status, the most-talked-about COVID-19-related symptoms we found (e.g., fatigue, cough, fever, headache, and difficulty breathing) were among the most common symptoms reported by COVID-19 patients in other studies.[33-35] Based on information extracted by applying COVID-19 disease status and diagnostic methods filters, we found that nearly 40% of social media users who discussed the top 5 most commonly mentioned symptom topics, such as fatigue and cough, also talked about the topic of having tested positive with COVID-19. We also noticed that about 15% of these discussions were related to asymptomatic COVID-19. While an in-depth exploration of these posts using qualitative analysis or sentiment analysis is necessary to help verify the users' COVID-19 disease status, our preliminary data indicate the potential for extracting information from social media to understand the full spectrum of symptoms experienced by COVID-19 patients. Interestingly, we noticed an increase of over 60% in the volume of posts mentioning less common COVID-19 symptoms such as changes in the senses of taste and smell during the second stage of our study period (September 1 to December 13, 2020). This surge may be partly due to improvements in knowledge and awareness of COVID-19 symptoms in the general population as the two symptoms were recently added to the CDC and WHO COVID-19 symptoms lists (late April and early May 2020, respectively).

While there have been fluctuations in the volume of social media posts from day-to-day, there appeared to be seasonal variation in the volume of discussion of symptoms and medical conditions. We noticed that the volume of most health-related discussions increased more from September 1 to December 13, 2020, than from June 14 to August 31, 2020. These changes may have been due to a combination of colder weather in the northern hemisphere and social distancing and limitations on daily life during the pandemic as well as the second wave of COVID-19, resulting in more social media users and more people being restricted indoors.[36] Additionally, there were several inflection points in the volume of discussion of symptoms and medical conditions in the last six months. These changes appeared to have coincided with major news stories and national events, echoing findings from other studies that showed the potential impact of media coverage on online discussions.[6, 28] For example, the volume of all five commonly mentioned symptoms (anxiety, generalized pain, weight loss, fatigue, and cough) and two medical conditions (unspecified CNS and depression) peaked on October 10, 2020, the day on which hurricane Delta struck Louisiana and nearby states and left 730,000 homes and businesses without power.[37] However, our study did not find evidence of an association between changes in the volume of symptom discussion over time and the trend of daily new confirmed cases of COVID-19 in the U.S.

Limitations

Our study is subject to several limitations. First, information on geolocation, demographics, and COVID-19 disease status was not available for all social media users in the study, due to various legal limitations (such as General Data Protection Regulation-EU GDPR). This might have introduced a sampling bias if there were significant differences between social media users' characteristics in our project and the real world. However, by collaborating with social media analytics companies, we have maximized our ability to access thousands of social media data sources worldwide, thus minimizing the possibility of sampling bias. Additionally, the majority of social media users in our study were from the U.S. The findings, therefore, may not be generalizable in their application to users located in other countries. Further, we did not conduct formal statistical analyses beyond comparing the trends differences in frequency of health-related posts and COVID-19 new cases; thus, further testing is needed to confirm the associations between patterns of changes in symptom/medical condition posts and the fluctuations of COVID-19 statistics over time. Finally, we did not perform sentiment analysis or qualitative analysis in the study and did not verify whether authors who discussed COVID-19 related topics had COVID-19 themselves. We hope to accomplish and report this analysis in a future study. We also hope that other studies on social media's role in public health will replicate and validate our exploratory findings in non-Twitter social media platforms.

Conclusions

In this research, we classified online posts, collected from June 14 to December 13, 2020, according to discussions of symptoms and medical conditions. Neuropsychological symptoms such as anxiety were the most frequently mentioned symptom subcategory. And COVID-19 infection was the most commonly mentioned medical condition. Our analysis also showed that health-related discussions were greater from September 1 to December 13, 2020, than from June 14 to August 31, 2020, aligning with the increase in U.S. COVID-19 cases during the winter months. These preliminary findings show promise for real-time monitoring of social media posts to measure the mental health status of a population during a global public health crisis and to assess the public's main health needs that have not been captured or met by the existing health system. Future research may incorporate information from social media into predictive models for the detection of emerging infectious diseases.

Acknowledgements

The authors thank Center for Outcomes Research and Evaluation Yale New Haven Hospital for their coordination of the project. Alina Cohen and Pini Matzner 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. Qinglan Ding takes full responsibility for the data interpretation and writing. All authors contributed to the editing and the approving of the final version of the paper for publication. 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

API: application programming interface

CDC: Centers for Disease Control and Prevention

COVID-19: coronavirus disease 2019

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

IP address: Internet Protocol address

NLP: natural language processing

U.K.: United Kingdom

U.S.: United States

WHO: World Health Organization

References

1. Dong, E., H. Du, and L. Gardner, *An interactive web-based dashboard to track COVID-19 in real time.* Lancet Infect Dis, 2020. **20**(5): p. 533-534.
2. Merchant, R.M. and N. Lurie, *Social Media and Emergency Preparedness in Response to Novel Coronavirus.* JAMA, 2020.
3. Clement, J. *Estimated U.S. social media usage increase due to coronavirus home isolation 2020*. 2020 [cited 2020 August 20].
4. Yousuf, H., et al., *Association of a Public Health Campaign About Coronavirus Disease 2019 Promoted by News Media and a Social Influencer With Self-reported Personal Hygiene and Physical Distancing in the Netherlands.* JAMA Netw Open, 2020. **3**(7): p. e2014323.
5. Abd-Alrazaq, A., et al., *Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study.* J Med Internet Res, 2020. **22**(4): p. e19016.
6. Wahbeh, A., et al., *Mining Physicians' Opinions on Social Media to Obtain Insights Into COVID-19: Mixed Methods Analysis.* JMIR Public Health Surveill, 2020. **6**(2): p. e19276.
7. Calvo, R.A., S. Deterding, and R.M. Ryan, *Health surveillance during covid-19 pandemic.* Bmj, 2020. **369**: p. m1373.
8. Li, H.O., et al., *YouTube as a source of information on COVID-19: a pandemic of misinformation?* BMJ Glob Health, 2020. **5**(5).
9. Merchant, R.M., et al., *Evaluating the predictability of medical conditions from social media posts.* PLoS One, 2019. **14**(6): p. e0215476.
10. Guntuku, S.C., et al., *Tracking Mental Health and Symptom Mentions on Twitter During*

COVID-19. J Gen Intern Med, 2020. **35**(9): p. 2798-2800.

11. Guntuku, S.C., et al., *Tracking Mental Health and Symptom Mentions on Twitter During COVID-19.* J Gen Intern Med, 2020: p. 1-3.

12. Kolliakou, A., et al., *Mental health-related conversations on social media and crisis episodes: a time-series regression analysis.* Sci Rep, 2020. **10**(1): p. 1342.

13. Fishman, J.M. and D. Casarett, *Mass media and medicine: when the most trusted media mislead.* Mayo Clin Proc, 2006. **81**(3): p. 291-3.

14. Czeisler, M., et al., *Delay or Avoidance of Medical Care Because of COVID-19–Related Concerns — United States, June 2020.* MMWR. Morbidity and Mortality Weekly Report, 2020. **69**.

15. Guan, W.J., et al., *Clinical Characteristics of Coronavirus Disease 2019 in China.* N Engl J Med, 2020. **382**(18): p. 1708-1720.

16. Young, B.E., et al., *Epidemiologic Features and Clinical Course of Patients Infected With SARS-CoV-2 in Singapore.* Jama, 2020. **323**(15): p. 1488-94.

17. K., S. and E. H. *Predicting COVID-19 Infection Groups using Social Networks and Machine Learning Algorithms*. in *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. 2020.

18. Merchant, R.M., *Evaluating the Potential Role of Social Media in Preventive Health Care.* Jama, 2020.

19. Bernstein, M., et al., *4chan and/b: An analysis of anonymity and ephemerality in a large online community.* 2011

20. Weichselbraun, A., et al., *Harvest -- An open source toolkit for extracting posts and post metadata from web forums*. 2021.

21. Tsou, M.H., H. Zhang, and C.T. Jung, *Identifying Data Noises, User Biases, and System Errors in Geo-tagged Twitter Messages (Tweets).* 2017.

22. Chu, Z., et al., *Who is tweeting on Twitter: human, bot, or cyborg?*, in *Proceedings of the 26th Annual Computer Security Applications Conference*. 2010, Association for Computing Machinery: Austin, Texas, USA. p. 21–30.

23. Signals Analytics. 2020: https://www.signals-analytics.com.

24. NetBase Quid. 2020: https://netbasequid.com.

25. Axisbits, *COVID-19 Coronavirus Statistics API Documentation*. 2020 p. https://rapidapi.com/axisbits-axisbits-default/api/covid-19-statistics/details.

26. Matzner, P., *Using advanced analytics for the early detection of pandemics and outbreaks.* . 2020 Signals Analytics: https://www.signals-analytics.com/resources/white-papers/early-detection-pandemics-outbreaks.

27. Centers for Disease Control and Prevention. *Symptoms of Coronavirus*. 2020 December 22, 2020 [cited 2020 December 22]; Available from: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html.

28. Valdez, D., et al., *Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data.* J Med Internet Res, 2020. **22**(12): p. e21418.

29. Ge, F., et al., *How to deal with the negative psychological impact of COVID-19 for people who pay attention to anxiety and depression.* Precision Clinical Medicine, 2020. **3**(3): p. 161-168.

30. Qiu, J., et al., *A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: implications and policy recommendations.* Gen Psychiatr, 2020.

33(2): p. e100213.

31. Zhou, L., et al., *Harnessing social media for health information management* Electron Commer Res Appl, 2018. **27**: p. 139-151.

32. Picone, M., et al., *Social Listening as a Rapid Approach to Collecting and Analyzing COVID-19 Symptoms and Disease Natural Histories Reported by Large Numbers of Individuals.* Popul Health Manag, 2020. **23**(5): p. 350-360.

33. Sarker, A., et al., *Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource.* J Am Med Inform Assoc, 2020.

34. Burke, R.M., et al., *Symptom Profiles of a Convenience Sample of Patients with COVID-19 - United States, January-April 2020.* MMWR Morb Mortal Wkly Rep, 2020. **69**(28): p. 904-908.

35. Alimohamadi, Y., et al., *Determine the most common clinical symptoms in COVID-19 patients: a systematic review and meta-analysis.* J Prev Med Hyg, 2020. **61**(3): p. E304-e312.

36. Merchant, R.M. and N. Lurie, *Social Media and Emergency Preparedness in Response to Novel Coronavirus.* JAMA, 2020. **323**(20): p. 2011-2012.

37. Aretakis, R. and G. Hauck, *Delta lives updates: Hundreds of thousands without power across south; Louisiana governor urges caution as clean-up begins*, in *USA Today*. 2020, Gannette Satellite information Network, LLC.

Supplementary Files

Multimedia Appendixes

List of contents Supplement 1. Developing data extraction query Supplement 2. Exclusions of advertisements and pornographic posts Supplement 3. Developing the taxonomy Supplement 4. Validating the taxonomy to demonstrate its utility Supplement 5. Applying the taxonomy (NLP algorithms) to extract posts mentioning symptoms and medical conditions topics automatically Supplement Figure 1a. “RBG” classified mentions in social media posts (September-October 2020) Supplement Figure 1b. “George Floyd” classified mentions in social media posts (June-July 2020) Supplemental Figure 2. The 5 most commonly mentioned symptoms on social media during the period of June 14 to December 13, 2020. Supplement Figure 3: The 5 most commonly mentioned medical disorders on social media during the period of June 14 to December 13, 2020. Supplement Figure 4. Comparing changes in the 5 most commonly mentioned symptoms on social media between June 13-Aug 31, 2020 to September 1-December 13, 2020. Supplement Figure 5. Comparing changes in the 5 most commonly mentioned medical disorders on social media between June 13-Aug 31, 2020 to September 1-December 13, 2020. Supplement Figure 6. Comparing changes in the 10 most commonly mentioned COVID-19 symptoms between June 13-Aug 31, 2020 and September 1-December 13, 2020. Supplement Table 1. Peak and trough dates of the number of social media posts for the 5 most commonly mentioned symptoms and medical conditions (June 14 to December 13, 2020).
URL: http://asset.jmir.pub/assets/2d9a1c0dd2d5630de60a866089239afb.docx