Self-reported COVID-19 symptoms on Twitter: An analysis and a research resource

Corresponding Author:

Abeed Sarker, PhD
101 Woodruff Circle
Office 4101
Atlanta, GA 30322
abeed@dbmi.emory.edu;
Phone: +1(602) 474 6203

Author List:

Abeed Sarker[1]

Sahithi Lakamana[1]

Whitney Hogg-Bremer[1]

Angel Xie[1]

Mohammed Ali Al-Garadi[1]

Yuan-Chi Yang[1]

[1]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, U.S.A

Keywords (MeSH)

Social media (MeSH ID: D061108)

Communicable Diseases (MeSH ID: D003141)

Virus Diseases (MeSH ID: D014777)

Natural language processing (MeSH ID: D009323)

Text mining (MeSH ID: D057225)

Article Type: Brief Communications

Word count:

Abstract: 149 words

Body: 2,000 words

Abstract

Objective

To mine Twitter to quantitatively analyze COVID-19 symptoms self-reported by users, compare symptom distributions against clinical studies, and create a symptom lexicon for the research community.

Materials and methods

We retrieved tweets using COVID-19-related keywords, and performed several layers of semi-automatic filtering to curate self-reports of positive-tested users. We extracted COVID-19-related symptoms mentioned by the users, mapped them to standard IDs, and compared the distributions with multiple studies conducted in clinical settings.

Results

We identified 203 positive-tested users who reported 932 symptoms using 598 unique expressions. The most frequently-reported symptoms were fever/pyrexia (65%), cough (56%), body aches/pain (40%), headache (35%), fatigue (35%), and dyspnea (34%) amongst users who reported at least 1 symptom. Mild symptoms, such as anosmia (26%) and ageusia (24%) were frequently reported on Twitter, but not in clinical studies.

Conclusion

The spectrum of COVID-19 symptoms identified from Twitter may complement those identified in clinical settings.

INTRODUCTION

The outbreak of the coronavirus disease 2019 (COVID-19) is now considered to be a global pandemic[1] and is estimated to be one of the worst pandemics in the known World history.[2] At the time of the writing of this article, over 2 million confirmed positive cases have been reported globally, with over 125,000 deaths.[3] As the pandemic continues to ravage the world, there is a flurry of research activities are being conducted whose focuses range from trialing possible vaccines and predicting the trajectory of the outbreak to exploring the characteristics of the virus better by studying those infected.

Studies focused on identifying the symptoms experienced by those infected by the virus have typically been conducted through patients who were hospitalized or received clinical care.[4–6] Many infected people only experience mild symptoms or are asymptomatic and do not seek clinical care, although the specific portion of asymptomatic carriers is unknown.[7–9] To better understand the full spectrum of symptoms experienced by infected people, there is a need to look beyond hospital- or clinic-focused studies. With this in mind, we explored the possibility of using social media, namely Twitter, to study symptoms self-reported by users who tested positive for COVID-19. Our primary goals were to (i) verify that users report their experiences with COVID-19—including their positive test results and symptoms experienced—on Twitter, and to (ii) compare the distribution self-reported of symptoms with those reported in studies conducted in clinical settings. Our secondary objectives were to (i) create a COVID-19 symptom corpus that captures the multitude of ways in which users express symptoms so that natural language processing (NLP) systems may be developed for automated symptom detection, and (ii) collect a cohort of COVID-19-positive Twitter users whose longitudinal self-reported information may be studied in the future. To the best of our knowledge, this is the first study that focuses on extracting COVID-19 symptoms from public social media. We will make the symptom corpus available with this paper to assist the research community, and it will be part of a larger, maintained data resource—a social media COVID-19 Data Bundle (https://sarkerlab.org/covid_sm_data_bundle/).

MATERIALS AND METHODS

Data collection and user selection

We collected data from Twitter via the twitter public streaming application programming interface (API). First, we used a set of keywords/phrases related to the coronavirus to detect tweets through the API: *covid*, *covid19*, *covid-19*, *coronavirus*, and *corona AND virus*, including their hashtag equivalents (*e.g.*, *#covid19*). Due to the high global interest in this topic, these keywords retrieved very large numbers of tweets. Therefore, we applied a first level of filtering to only keep tweets that also mentioned at least one of the following terms: *positive*, *negative*, *test* and *tested*, along with one of the personal pronouns *I*, *my*, *us*, *we*,

and *me*, and only these tweets were stored in our database. To discover users who self-reported positive COVID-19 tests, we applied a third layer of filtering using regular expressions. We used the expressions '*i.*test[ed] positive*', '*we.*test[ed] positive*', '*test.*came back positive*', '*my.*[covid|coronavirus|covid19].*symptoms*', '*[covid|coronavirus|covid19].*[test|tested].*us*'. We also collected tweets from a publicly available coronavirus dataset,[10] and applied the same layers of filers. This provided us with a more manageable set of manual review, although most were still false positives (*e.g.*, '*... I dreamt that I tested positive for covid ...*'). We manually reviewed the tweets to identify true self-reports, while discarding the clear false positives. We found some users falsely reporting positive tests, copying posts by celebrities or initially claiming positive tests only to clarify in later tweets that the tests actually came back negative, and we removed these users from our COVID-19-positive set. These multiple layers of filtering gave us a manageable set of potential COVID-19-positive users whose tweets we could analyze semi-automatically. The filtering decisions were made on iteratively—by collecting sample data for hours and days, and then updating the collection strategy based on analyses of the collected data.

Symptom discovery from user posts

For all the COVID-19-positive users identified, we collected all their past posts dating back to February 1, 2020. We did not include tweets earlier than that because we assumed that symptoms posted prior to that would not be related to COVID-19, particularly because our data collection started in late February and most of the positive test announcements we detected were from late March to early April. Since we were interested in identifying patient-reported symptoms only in this study, we tried to shortlist tweets that were likely to mention symptoms. To perform this, we first created a meta-lexicon by combining MedDRA,[11] Consumer Health Vocabulary (CHV),[12] and SIDER.[13] Lexicon-based approaches are known to have low recall particularly for social media data since social media expressions are often non-standard and contain misspellings.[14,15] Therefore, instead of searching the tweets for exact expressions from the tweets, we performed inexact matching using a string similarity metric. Specifically, for every symptom in the lexicon, we searched windows of sequences of characters in each tweet that had similarity values above a specific threshold. We used the levenshtein ratio as the similarity metric, computed as $1 - \frac{lev.\ dist.}{\max(length)}$, where *lev. dist.* represents the levenshtein distance between the two strings and *max(length)* represents the length of the longer string. Our intent was to attain high recall, so that we were unlikely to miss possible expressions of symptoms while also filtering out many tweets that were completely off topic. We set the threshold via trial and error over sample tweets, and because of the focus on high recall, this approach still retrieved many false positives (*e.g.*, tweets mentioning body parts but not in the context of an illness or a symptom). After running this inexact matching approach on approximately 50 user profiles, we manually extracted the true positive expressions (*i.e.*, those that expressed symptoms in the context of a COVID-19) and added them

to the existing lexicons. We also manually reviewed all the tweets from these profiles to ensure that our similarity-matching approach did not miss any possible symptoms.

Following these multiple filtering methods, we manually reviewed all the posts from all the users, identified each true symptom expressed, and removed the false positives. We semi-automatically mapped the expressions to standardized concept IDs in the Unified Medical Language System (UMLS) using the meta-lexicon we developed and the NCBO BioPortal.[16] In the absence of exact matches, we searched the BioPortal to find the most appropriate mappings. For user profiles that had too few or no tweets remaining after the filtering processes, we manually checked their Twitter profiles using the web interface. All annotations and mappings were reviewed, reviewers' questions were discussed at meetings. In general, we found that it was easy for annotators to detect expressions of symptoms, even when the expressions were non-standard ('pounding in my head' = Headache).

Following the mapping process, we computed the frequencies of the patient-reported symptoms on Twitter and compared them with several other recent studies that used data from other sources. We also identified users who reported that they had tested positive and also specifically stated that they showed '*no symptoms*'. When computing the frequencies and percentages of symptoms, we used two models: (i) computing raw frequencies over all the detected users, and (ii) computing frequencies for only those users who reported at least 1 symptom or explicitly stated that they had no symptoms. We believe the frequency distribution for (ii) was more robust as for users who reported no specific symptoms, we could not verify if they had actually experienced any symptoms and not reported them or just did not share symptoms over Twitter.

RESULTS

Our initial keyword-based data collection and filtering from the different sources retrieved millions of tweets, excluding retweets, and we found repeated tweets to be a major problem. This resulted from celebrities or verified users with many followers posting about their own experiences with COVID-19, and then many users re-posting (not retweeting) the exact texts. Removing such duplicate tweets left us with 305 users (472,018 tweets) to review. 102 of them were labeled as '*negatives*'—users who had specifically stated that their tests had come back negative, who removed their original posts about testing positive, or whose positive-test-indicating posts appeared to be fake or inconclusive after a detailed review of the account (*e.g.*, we found some users claiming they tested positive as an April fool joke). This left us with 203 true positive users, with 117,572 tweets, whose positive testing posts were deemed to be legitimate. The symptom detection approach reduced the number of unique tweets to review to 6,438.

The 203 users expressed 932 total symptoms (mean: 4.59 per user; median: 4) using 598 unique expressions, which we grouped into 48 categories (Table 1). We computed two sets of percentages—for all users (n=203) and for users who expressed at least 1 symptom or stated that they were asymptomatic (n=169). 34 users did not did not express any specific symptoms, while 10 users explicitly mentioned that they experienced no symptoms. As can be seen from the table, *fever/pyrexia* was the most commonly reported symptom, followed by *cough*, *body ache & pain*, *headache*, *fatigue*, *dyspnea*, *chills*, *anosmia*, *ageusia*, *throat pain and chest pain*, with each mentioned by 20% of the users who reported at least one symptom. Figure 1 illustrates the distribution of the number of symptoms reported by the users in our cohort.

Table 1. *Distribution of symptoms reported by COVD-19 positive users on Twitter. Symptoms expressed once only are grouped as "Other symptoms".*

Symptom	**Raw count**	**Percentage All**	**Percentage for > 0**
Pyrexia	109	54%	65%
Cough	94	46%	56%
Body ache & Pain	67	33%	40%
Headache	59	29%	35%
Fatigue	59	29%	35%
Dyspnea	57	28%	34%
Anosmia	44	22%	26%
Chills	41	20%	24%
Ageusia	41	20%	24%
Oropharyngeal pain	39	19%	23%
Chest pain	33	16%	20%
Chest tightness	22	11%	13%
Sweats	20	10%	12%
Loss of appetite	20	10%	12%
Nausea	17	8%	10%
Rhinorrhea	16	8%	10%
Vomit	16	8%	10%
Anxiety, stress & general mental health sympoms	16	8%	10%
Pneumonia	15	7%	9%

Migraine	13	6%	8%
Diarrhea/GI issues	12	6%	7%
Eye pain	10	5%	6%
Dizziness/disorientation/confusion	10	5%	6%
No Symptoms	10	5%	6%
General experiences	9	4%	5%
Lethargic	9	4%	5%
Myalgia	9	4%	5%
Sneezing	8	4%	5%
Insomnia/sleep disturbance	8	4%	5%
Paranasal sinus discomfort	7	3%	4%
Upper respiratory tract infection	6	3%	4%
Wheezing	5	3%	3%
Ear infection/pain	5	3%	3%
Dehydration	4	2%	2%
Palpitations	4	2%	2%
Abdominal pain	3	2%	2%
Hot Flush	2	1%	1%
Arthralgia	2	1%	1%
Nasal dryness	2	1%	1%
Other symptoms	8	4%	5%

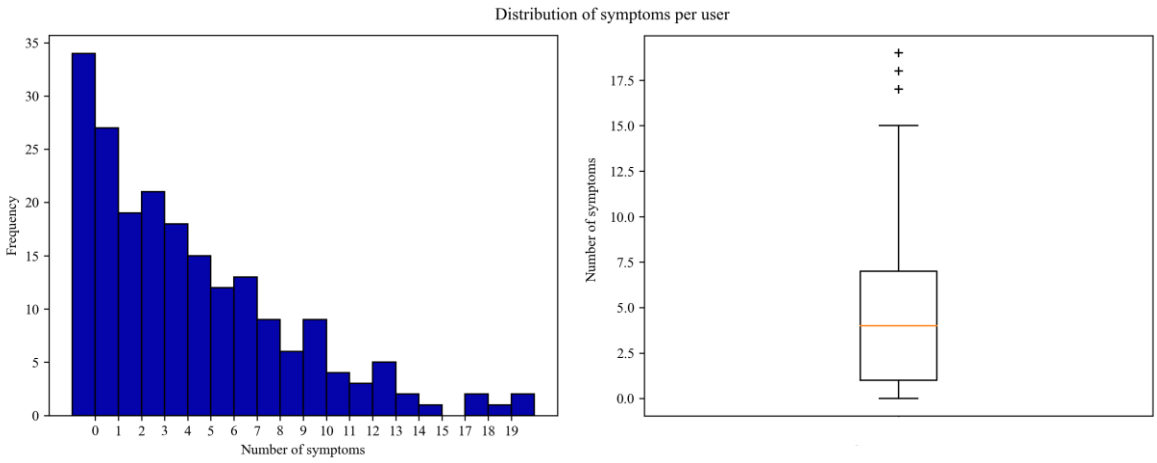

Figure 1. *Distribution of the number of symptoms reported by our Twitter cohort.*

Table 2. *Comparison of common symptoms reported on Twitter with those reported in clinical settings.*

Symptom	**Our study**[θ] (n=169) N (%)	**Huang et al.**[6] (n=41) N (%)	**Chen et al.**[5] (n=249) N (%)	**Wang et al.**[17] (n=138) N (%)	**Chen et al.**[18] (n=99) N (%)	**Guan et al.**[4] (n=1099) N (%)	**WHO Report**[19] (n=55,924) N (%)
Fever (Pyrexia)	109 (65)	40 (98)	217 (87)	136 (99)	82 (83)	975 (89)	49157 (88)
Cough	94 (46)	31 (76)	91 (37)	82 (59)	81 (82)	745 (68)	37861 (68)**
Dyspnea	57 (34)	22/40 (55)	19 (8)	43 (31)	31 (31)	205 (19)	10402 (19)
Headache	59 (35)	3/38 (8)	28 (11)[∀]	9 (7)	8 (8)	150 (14)	7606 (14)
Body ache	67 (40)	-	-	48 (35)[Ψ]	11 (11)[Ψ]	164 (15)[Ψ]	8277 (15)[Ψ]
Fatigue	59 (35)	18 (44)*	39 (16)	96 (70)	-	419 (38)	21307 (38)
Chills	41 (24)	-	-	-	-	126 (12)	6375 (11)
Anosmia	44 (26)	-	-	-	-	-	-
Ageusia	41 (24)	-	-	-	-	-	-
Chest pain	33 (20)	-	-	-	2 (2)		-
Oropharengeal pain (sore throat)	39 (23)	-	16 (6)	24 (17)	5 (5)	153 (14)	7773 (14)
Diarrhea	12 (7)	1/38 (3)	8 (3)	14 (10)	2 (2)	42 (4)	2069 (4)
Rhinorrhea	16 (10)	-	17 (7)	-	4 (4)	53 (5)	2684 (5)[ΨΨ]
Anorexia	20 (12)	-	8 (3)	55 (40)	-	-	-
Nausea	17 (10)	-	-	14 (10)	1 (1)[♠]	55 (5)[♠]	2796 (5)[♠]
Asymptomatic	10 (6)	-	7 (3)	-	-	-	-

[θ]For users who expressed at least 1 symptom or expressed that they did not have any symptoms.
*The study provided a combined number for myalgia and fatigue.
[∀]Headache and dizziness was combined for this study.
[Ψ]The reported number is for myalgia/muscle ache and/or arthralgia. In our study, we separated myalgia, arthralgia, body ache and pain.
[♠]Nausea and vomiting as a single category.
**Reported as dry cough.
[ΨΨ]Reported as nasal congestion.

Table 2 compares the symptom percentages reported by our Twitter cohort with several studies conducted in clinical settings (*i.e.*, patients who were either hospitalized or visited hospitals/clinics for treatment). The top symptoms remained fairly consistent across the studies—*fever/pyrexia*, *cough*, *dyspnea*, *headache*, *body ache* and *fatigue*. The percentage of fever (65%), though the highest in our dataset, is lower than all the studies conducted in clinical settings. In our study, we distinguished, where possible, between *myalgia* and *arthralgia*, and combined *pain (general)* and *body ache*. Combining all these into one category, as some studies had done, would result in a higher proportion. We found considerable numbers of reports of *anosmia* (26%) and *ageusia* (24%), with approximately one-fourth of our cohort reporting these symptoms. Reports of these symptoms, however, were missing from the other studies.

DISCUSSION AND CONCLUSIONS

Our study revealed that there were many self-reports of COVID-19 positive tests on Twitter, although such reports are buried in large amounts of noise. We observed a common trend among Twitter users of describing their day-to-day disease progression since the onset of symptoms. This trend perhaps became popular as celebrities started describing their symptoms on Twitter. We saw many reports from users who tested positive but initially showed no symptoms, and some who expressed anosmia and/or ageusia as the only symptoms, which were documented in the comparison studies. There are some studies that suggest that anosmia and ageusia may be the only symptom of COVID-19 among otherwise asymptomatic patients.[20–22] The most likely explanation behind the differences between symptoms reported on Twitter versus clinical studies is that the former were reported mostly by users who had milder infections, while people who visited hospitals often went there to receive treatment for more serious symptoms. Also, the median ages of the patients studied in clinical studies tended to be much higher than the median age of Twitter users (in the U.S., median Twitter user age is 40[23]). In contrast to the clinical studies, in our cohort, some users expressed stress and anxiety suffered after testing positive. It was difficult in many cases to ascertain if the mental health issues were directly related to COVID-19 or whether the users had prior histories of such conditions.

To the best of our knowledge, this is the first study to have utilized Twitter to curate symptoms posted by COVID-19-positive users. In the interest of community-driven research, we will be making the symptom lexicon available with this publication. The cohort of users detected over social media will enable us to conduct further studies in the future, enable us to study relatively unexplored topics such as the mental health impacts of the pandemic, and the long-term health of those infected by the virus.

FUNDING

TBA

COMPETING INTERESTS

None declared.

CONTRIBUTIONS

AS designed the study and data collection/filtering strategies. All authors contributed to the analyses, annotation process, and the writing of the manuscript.

ACKNOWLEDGMENTS

None.

REFERENCES

1. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Accessed April 12, 2020.

2. Sansa NA. Effects of the COVID-19 Pandemic on the World Population: Lessons to Adopt from Past Years Global Pandemics. *SSRN Electron J*. April 2020. doi:10.2139/ssrn.3565645

3. COVID-19 Map - Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html. Accessed April 12, 2020.

4. Guan W, Ni Z, Hu Y, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med*. February 2020. doi:10.1056/nejmoa2002032

5. Chen J, Qi T, Liu L, et al. Clinical progression of patients with COVID-19 in Shanghai, China. *J Infect*. March 2020. doi:10.1016/j.jinf.2020.03.004

6. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395(10223):497-506. doi:10.1016/S0140-6736(20)30183-5

7. Bai Y, Yao L, Wei T, et al. Presumed Asymptomatic Carrier Transmission of COVID-19. *JAMA - J Am Med Assoc*. 2020. doi:10.1001/jama.2020.2565

8. Chan JFW, Yuan S, Kok KH, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*. 2020;395(10223):514-523. doi:10.1016/S0140-6736(20)30154-9

9. Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N Engl J Med*. 2020;382(13):1199-1207. doi:10.1056/NEJMoa2001316

10. Chen E, Lerman K, Ferrara E. COVID-19: The First Public Coronavirus Twitter Dataset. March 2020. http://arxiv.org/abs/2003.07372. Accessed April 12, 2020.

11. Mozzicato P. MedDRA: An overview of the medical dictionary for regulatory activities.

Pharmaceut Med. 2009;23(2):65-75. doi:10.1007/BF03256752

12. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. *J Am Med Informatics Assoc*. 2006. doi:10.1197/jamia.M1761

13. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res*. 2015;44:1075-1079. doi:10.1093/nar/gkv1075

14. Yazdavar AH, Al-Olimat HS, Ebrahimi M, et al. Semi-Supervised approach to monitoring clinical depressive symptoms in social media. In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*. New York, New York, USA: Association for Computing Machinery, Inc; 2017:1191-1198. doi:10.1145/3110025.3123028

15. Zhang L, Ghosh R, Dekhil M, Hsu M, Liu B. *Combining Lexicon-Based and Learning-Based Methods for Twitter Sentiment Analysis*.; 2011.

16. Welcome to the NCBO BioPortal | NCBO BioPortal. https://bioportal.bioontology.org/. Accessed April 15, 2020.

17. Wang D, Hu B, Hu C, et al. Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA - J Am Med Assoc*. 2020;323(11):1061-1069. doi:10.1001/jama.2020.1585

18. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet*. 2020;395(10223):507-513. doi:10.1016/S0140-6736(20)30211-7

19. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). https://www.who.int/publications-detail/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19). Accessed April 15, 2020.

20. Gane SB, Kelly C, Hopkins C. Isolated sudden onset anosmia in COVID-19 infection. A novel syndrome? *Rhinology*. April 2020. doi:10.4193/Rhin20.114

21. Hjelmesæth J, Skaare D. Loss of smell or taste as the only symptom of COVID-19. *Tidsskr Den Nor legeforening*. April 2020. doi:10.4045/TIDSSKR.20.0287

22. Villalba NL, Maouche Y, Ortiz MBA, et al. Anosmia and Dysgeusia in the Absence of Other Respiratory Diseases: Should COVID-19 Infection Be Considered? *Eur J Case Reports Intern Med*. April 2020. doi:10.12890/2020_001641

23. How Twitter Users Compare to the General Public | Pew Research Center. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/. Accessed April 12, 2020.