AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

OXFORD

Brief Communications

Mining long-COVID symptoms from Reddit: characterizing post-COVID syndrome from patient reports

Abeed Sarker [iD] and Yao Ge

Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, Georgia, USA

Corresponding Author: Abeed Sarker, PhD, Department of Biomedical Informatics, School of Medicine, Emory University, 101 Woodruff Circle, Office 4101, Atlanta, GA 30322, USA; abeed@dbmi.emory.edu

ABSTRACT

Our objective was to mine Reddit to discover long-COVID symptoms self-reported by users, compare symptom distributions across studies, and create a symptom lexicon. We retrieved posts from the */r/covidlonghaulers* subreddit and extracted symptoms via approximate matching using an expanded meta-lexicon. We mapped the extracted symptoms to standard concept IDs, compared their distributions with those reported in recent literature and analyzed their distributions over time. From 42 995 posts by 4249 users, we identified 1744 users who expressed at least 1 symptom. The most frequently reported long-COVID symptoms were *mental health-related symptoms* (55.2%), *fatigue* (51.2%), *general ache/pain* (48.4%), *brain fog/confusion* (32.8%), and *dyspnea* (28.9%) among users reporting at least 1 symptom. Comparison with recent literature revealed a large variance in reported symptoms across studies. Temporal analysis showed several persistent symptoms up to 15 months after infection. The spectrum of symptoms identified from Reddit may provide early insights about long-COVID.

Key words: social media, COVID-19, post-acute COVID-19 syndrome, virus diseases, natural language processing

---

LAY SUMMARY

Many people continue to experience debilitating symptoms long after recovering from their initial COVID-19 infection. This syndrome is commonly referred to as "long-COVID" and the sufferers are called "COVID long haulers." The phrase *long-COVID* was coined by *long haulers*, who organized online after continuing to experience persistent long-term symptoms, and raised awareness until healthcare professionals acknowledged its existence. With many unknowns associated with this syndrome, social media is a potentially valuable resource to better understand it as many people discuss their long-COVID experiences on it. In this article, we applied natural language processing to study symptoms associated with long-COVID that are reported by users on the *r/covidlonghaulers* subreddit of Reddit. We found that the most commonly reported long-COVID symptoms are *mental health-related symptoms, fatigue, general/body ache and pain, confusion/disorientation/brain fog*, and *dyspnea* (shortness of breath). An analysis of symptom distribution over time showed that some patients continue to experience symptoms up to 15 months following their initial diagnosis, with symptoms such as *fatigue* being more persistent than others. We conclude that surveillance of social media data may help the medical community to better understand long-COVID from the perspective of patients and enable them to improve medical care.

INTRODUCTION

The persistence of symptoms 28 days after the diagnosis of SARS-CoV-2 (COVID-19) virus infection is referred to as "post-acute COVID syndrome," "post-acute sequelae of SARS-CoV-2," or "long-COVID."[1–3] A growing body of research is exploring long-COVID and the constellation of symptoms that patients experience following acute COVID-19 infection.[3–6] The phrase "long-COVID" was coined by patients ("long haulers"), who organized online, started a grass-roots movement after continuing to experience persistent long-term symptoms, and raised awareness until healthcare professionals acknowledged the existence of the condition.[7] Ongoing research suggests that long-COVID may affect people of varying ages and those with asymptomatic acute infections.[8–12] There are, however, many unknowns associated with long-COVID, including its prognosis, the distribution of symptoms, and their long-term manifestation. It is now well-established that long-COVID is a complex health problem affecting millions of people globally, and a wide range of symptoms are being discovered over time.[5,9,12,13] Consequently, this syndrome is emerging as a critical research topic relevant to the ongoing COVID-19 pandemic.

Since long-COVID awareness campaigns first started on social media, it is a potential source of early insights regarding the topic. Many people, including healthcare professionals who had COVID-19 infections, report their experiences through this medium.[14,15] Social media adoption is currently at an all-time high globally and in the United States,[16,17] and user-generated information from social media have been utilized to study health-related topics, including infectious diseases.[18,19] Due to the growing abundance of patient-reported knowledge about COVID-19 on social media, it has been used to study vaccine hesitancy,

misinformation, and related topics.[20–23] In this study, we mined publicly available social media data from Reddit to discover and analyze the spectrum of symptoms self-reported by COVID-19 long haulers.

Objectives and contributions

Our objectives for this study were to (1) extend a previously developed COVID-19 symptom lexicon,[23] (2) deploy the extended lexicon to identify self-reported long-COVID symptoms from a dedicated forum on Reddit, (3) analyze the distribution of symptoms, and (4) compare them with symptoms reported in recent studies. Our intended primary contribution is to add to the evolving body of knowledge about long-COVID based on data-driven methods and social media data. Our findings describe patient-reported symptom distributions and show that many long-COVID patients continue to report symptoms up to 15 months following acute infection. Additional, informatics-oriented contributions include the creation of a Reddit-specific long-COVID symptom lexicon with mappings to concept IDs in the unified medical language system (UMLS)[24] and a method for extracting symptoms automatically from free text. The lexicon and methods may potentially be used for conducting syndromic surveillance using social media.

MATERIALS AND METHODS

Data source and collection

We collected data from the subreddit /r/covidlonghaulers on Reddit, which is a moderated forum for discussing long-COVID-related topics with strict rules prohibiting the promotion of alternative treatments, misinformation, and conspiracy theories.[25] As of July 31

Table 1. Statistics associated with the dataset included in this study

Posts	Discussion threads	Unique users	Users reporting at least 1 non-negated symptom	Users reporting at least 1 symptom without accounting for negation
42 995	1220	4249	1744	2576
Total symptom expressions detected	Negated symptoms	Number of sentences	Number of words/tokens	Mean post length in words
47 700	1532	171 937	2 815 655	65.49
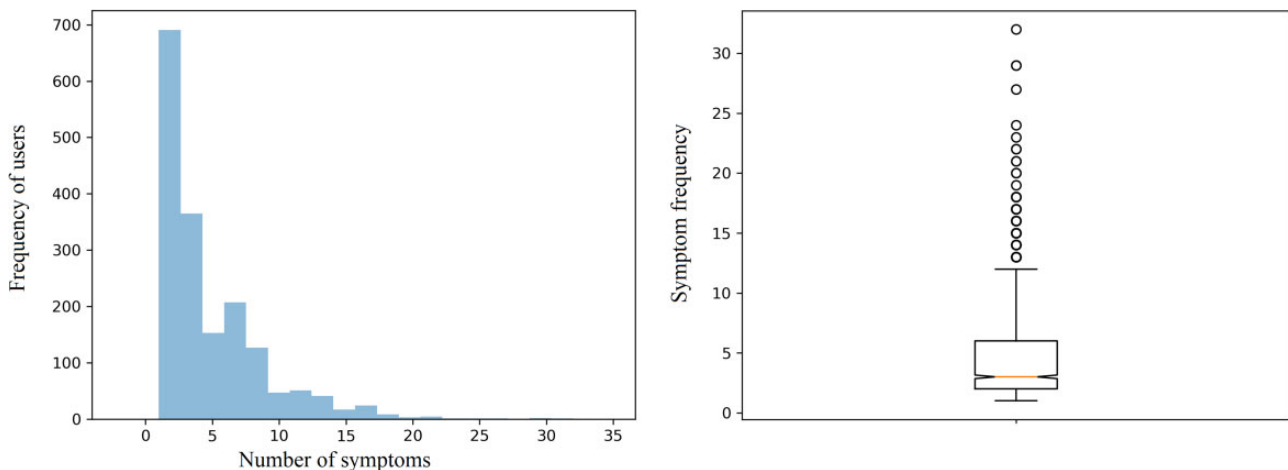
Figure 1. Distribution of numbers of long-COVID symptoms reported by Reddit users in the *r/covidlonghaulers* subreddit.
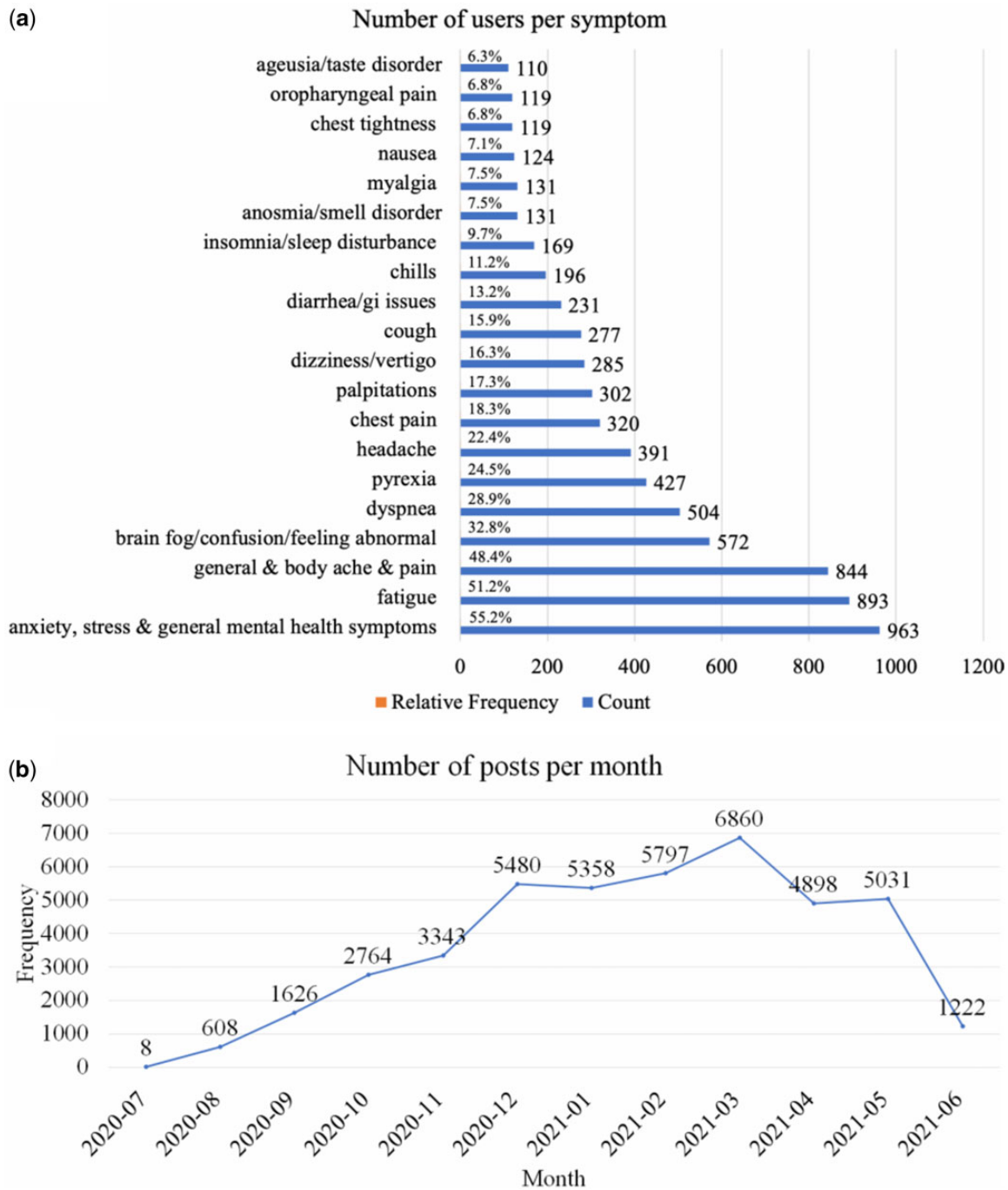
Figure 2. (A) Number of users reporting each of the top 20 symptoms and their relative frequencies to the total number of users. (B) Distribution of posts per month for the dataset included.

2021, the subreddit had over 16 000 subscribers and over 1400 threads—most focusing on self-reported long-COVID experiences. In the subreddit, users report and discuss the symptoms they experience and often provide timelines of their symptoms. It is thus an excellent resource for obtaining crowdsourced information about long-COVID syndrome. We collected all the available posts from this subreddit using the PRAW application programming interface (API).[26] The earliest post included in our analyses was from July 25 2020 and the latest post was from June 6 2021. The API enables the collection of all available past data from the subreddit and is not a real-time streaming API. Consequently, posts that violated the community rules and were subsequently removed by the moderators were not retrieved by the API.

Symptom extraction from user posts

We first manually reviewed posts from the subreddit and annotated a symptom lexicon, grouping similar symptoms, and mapping expressions to UMLS concept IDs using the National Center for Biomedical Ontology BioPortal.[27] We augmented the lexicon entries for rarely
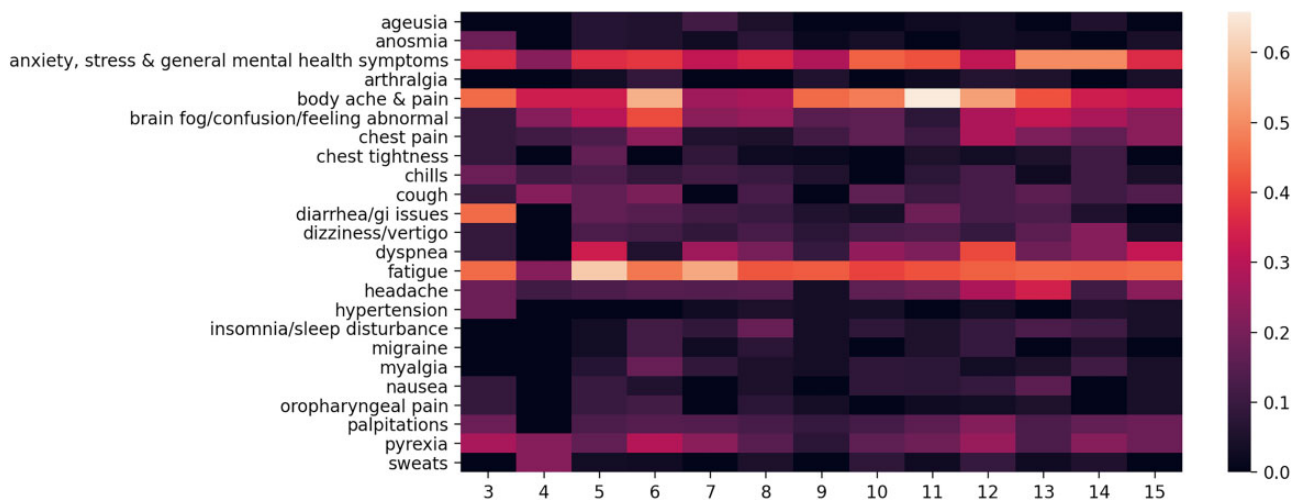
Figure 3. Heatmap of relative frequencies of symptoms from months 3 to 15.

occurring symptoms by adding expressions from the Medical Dictionary for Regulatory Activities.[28] We combined this lexicon with a previously developed Twitter lexicon[23] to create a meta-lexicon consisting of common COVID-19 and long-COVID symptom expressions. We further expanded the meta-lexicon automatically using the LexExp tool, which generates multi-word spelling variants.[29]

To detect symptoms from free text, we applied an inexact matching method. Exact matching on social media free texts typically results in low recall due to the presence of nonstandard expressions and misspellings. To overcome this issue, we searched through term sequences in the texts and computed the similarity of each prespecified finite word sequence with all entries in the meta-lexicon. Text sequences that obtained similarities above a specific threshold with any entry in the lexicon were considered to be candidates for long-COVID symptoms. We used the Levenshtein ratio metric to compute similarity, with a threshold of 0.98. For an expression of n words, the term sequences were of the range $[n - 1 : n + 2]$, similar to prior work.[23] During annotation, we also discovered that users often expressed the resolution or absence of symptoms using negation expressions such as "no," "do not," "never had," and we created a lexicon containing the negations we encountered. After identifying candidate symptoms, the symptom extraction algorithm determines if it might be negated by first checking if a negation expression encoded in the lexicon occurs in the post, and, if yes, if the candidate expression potentially falls within the scope of the negation expression. To detect potentially negated symptoms in this manner, the extraction algorithm preprocesses the texts by lowercasing and tokenizing, and labels any candidate symptom appearing within a 3-word window or before an end-of-sentence marker (eg, a period) following the negation term as negated. Finally, the symptoms are grouped per user and the frequency distributions for all users are generated. All parameters mentioned above were optimized via trial and error over the posts that were used for annotation.

We evaluated the performance of the algorithm using the precision, recall, and F_1-score metrics (F_1-score is the harmonic mean of precision and recall; F_1-score$= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$). To estimate precision and recall, we selected a random sample of posts (excluding those that were used for lexicon creation and system development), including posts from which at least 1 symptom was automatically detected and those with no detected symptoms. Negated symptoms extracted by the algorithm were counted as false positives.

Symptom distribution by month

Many thread titles in our dataset specified the number of months since the initial COVID-19 diagnosis that the particular user was experiencing symptoms for (eg, "8-month update," "11-month long hauler just got first vaccine"). We created a set of regular expressions to detect such expressions and computed the relative frequency distributions for the numbers of users expressing at least 1 symptom within these posts. We grouped the frequencies by month, starting from 3 up to 15.

RESULTS

Table 1 summarizes the statistics of the dataset included in this study. To evaluate algorithm performance, we manually reviewed posts associated with a total of 450 symptom expressions along with 25 randomly selected posts that had no detected symptoms. The approach obtained F_1-score of 0.94 (recall: 0.93; precision: 0.95). Figure 1 presents the distribution of the numbers of symptoms reported by this cohort. Most users reported 1–6 symptoms (histogram; left), and the median number of symptoms reported per user was 3 (boxplot; right). A handful of users reported 12 or more symptoms, with 32 being the highest number of symptoms expressed by a single user.

Figure 2A presents the raw counts for the top 20 symptoms reported by the users and their relative frequencies as a percentage of the total number of users who reported at least 1 non-negated symptom. There are some interesting differences between previously reported acute COVID-19 symptoms from Twitter[23] and the long-COVID symptoms we detected in this study. The most commonly reported long-COVID symptoms are *anxiety/stress and related mental health symptoms* (we grouped these symptoms into one because users often use non-standard expressions for them and it is difficult to separate such symptoms in a more fine-grained manner). The next 4 most frequently reported symptoms are *fatigue*, *general and body ache and pain*, *confusion/disorientation* (commonly expressed as *brain fog*), and *dyspnea*. *Pyrexia*, which was the most frequently reported symptom among an acute COVID-19 cohort from Twitter,

Table 2. Comparison of long-COVID symptom distributions (as percentages) identified in this study with those reported in recent literature for frequently reported symptoms

Study timeline	This study (n=1744) Up to 15 months post-acute infection	Osikomaiya et al[30] (n=274) 2 weeks post-discharge from isolation facilities	Huang et al[31] (n=1733) 6 months post-acute infection	Sykes et al[32] (n=134) 46–167 days post-discharge	Mandal et al[33] (n=276) Median of 54 (IQR 47–59) days post-discharge	Carfì et al[34] (n=143) Mean 60.3 (SD, 13.6) days post-onset of first symptom	Orrù et al[35] (N=152) >3 months post-positive COVID-19 test	Garrigues et al[36] (n=120) Mean 110.9 (SD, 11.1) days post-admission	Lopez-Leon et al[37] (n=N/A) Varying—meta-analysis of multiple studies
Fatigue	51.2	12.8	63	39.6	69	53.1	74.3	55.0	58
Ache and pain	48.4	8.8[a]	9[b]	51.5[a]	N/A	27.3[b]	61.2[a]	N/A[c]	19[b]
Brain fog/confusion or dizziness	32.8	5.2[d]	6[e]	25.4[d]	N/A	8[f,g]	48.7	26.7[d]	27[d]
Dyspnea	28.9	9.5	N/A	59.7[h]	53[h]	43.4	40.1	41.7	24
Fever/pyrexia	24.5	6.2	<1	10.4	N/A	N/A (excluded from study)	19.1	N/A	11[i]
Headache	22.4	12.8	2	N/A	N/A	10[f]	46.7	N/A	44
Insomnia/sleep disturbance	9.7	9.8	26	35.1	N/A	N/A	N/A[j]	30.8	11
Chest pain	18.3	9.8	5	17.9	N/A	21.7	26.3	10.8	16
Cough	15.9	9.2	N/A	35.1	34	18[f]	21.1	16.7	19
Palpitations	17.3	7.4	9	N/A	N/A	N/A	38.8	N/A	11

Note: The *Study timeline* row specifies the post-COVID follow-up period during which the information were collected.

Abbreviations: IQR: interquartile range; N/A: not available; SD: standard deviation.

[a]Reported symptom is *myalgia*.

[b]Only joint pain (*arthralgia*) reported.

[c]Patients reported pain/discomfort using a 5-point scale.

[d]Reported as attention or memory deficit or disorder.

[e]Only dizziness reported.

[f]Estimated from figure.

[g]Reported as *vertigo*.

[h]Reported as breathlessness.

[i]Reported as intermittent fever.

[j]Sleep disturbances were measured based on ISI scores.

is reported much less frequently for long-COVID (6th). Figure 2B shows the distribution of posts by month, illustrating there was a gradual increase in the number of posts every month until December 2020. Note that the number of posts in June 2021 only represents the first 6 days of the month.

The heatmap in Figure 3 presents the relative frequencies of 24 symptoms from months 3 to 15. The figure suggests that *general and body ache and pain, fatigue, anxiety, stress and other mental health issues,* and *dyspnea* were the most persistent symptoms, with a large number of users reporting to experience it 12 months after their initial diagnosis. Some symptoms, such as *pyrexia, GI issues,* and *oropharyngeal pain* appear to decrease over time. Table 2 presents the relative frequency distributions of 10 commonly reported long-COVID symptoms from this study and 8 recently published papers on long-COVID. *Fatigue* is consistently reported with high frequency across studies, but considerable variations can be observed in the symptom distributions.

DISCUSSION

Our findings broadly agree with recently published studies on long-COVID. The ranges for the symptom frequency distributions in Table 2 are wide, likely due to the differing patient populations included in the studies. Another possibility is that long-COVID experiences of patients vary considerably compared to the experiences of acute COVID-19 patients. Our findings also show that many patients continue to report symptoms up to 15 months after acute infection. Continuing surveillance of this cohort may reveal how the symptoms resolve, if at all. It is evident that even if new COVID-19 cases are lowered globally through vaccinations, an ongoing challenge will be to address complications associated with long-COVID.

With limited scientific data available about long-COVID and with global medical care primarily focusing on treating acute COVID-19 cases, many long-COVID sufferers are turning to social media to discuss their persistent symptoms, find others with similar symptoms, and identify potential solutions to improve quality of life. It is possible, and perhaps likely, that many patients will continue to suffer from long-COVID symptoms in the future, and social media will be an invaluable resource for obtaining early, crowd-sourced insights about the topic. While manually reviewing the posts, we found many users expressing their frustrations regarding their healthcare providers not understanding the extent of their sufferings and/or not being able to diagnose the underlying reasons behind their persistent symptoms. For example, many users who reported *dyspnea* stated that they had no issues with their lungs during the acute phases of their COVID-19 infections, and their care providers, who relied on standard imaging methods for diagnosis, dismissed their sufferings. We also found a number users explaining that their symptoms were not continuous, but cyclic—disappearing and returning from time to time. Users also discussed that their long-COVID symptoms were preventing them from getting back to their usual lives and work, adding to their mental health problems.

Our study has several limitations: (1) symptom extraction is not 100% accurate, (2) users may not report all their symptoms (leading to underestimation of prevalence), (3) Reddit users tend to be younger compared to the general population, and (4) rare symptoms may have been missed during lexicon preparation. Our choice of the data source, however, mitigates some common issues with social media data (eg, Twitter) such as misinformation, since the forum is heavily moderated. In the future, we will attempt to analyze how long-

COVID typically progresses, the distribution of time spans for symptom resolution, and treatment regimens that appear to help.

CONCLUSIONS

Social media data may help the medical community to better understand long-COVID from the perspective of patients and thus enable them to improve medical care. Considering the large volume of data that is generated on social media about this topic, it is necessary to develop automated methods involving NLP for long-term surveillance. To aid future research, we have made the lexicon developed for this study publicly available.

AUTHOR CONTRIBUTIONS

AS was responsible for the conception, design, and analyses presented in the manuscript. YG assisted with data preparation and evaluation of the methods. Both authors contributed to the writing of the final manuscript and are accountable for all aspects of the work.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY STATEMENT

The data used in this article was publicly available at the time of collection on the *r/covidlonghaulers* subreddit of Reddit (URL: https://www.reddit.com/r/covidlonghaulers/). The lexicon prepared by the authors will be made available via https://sarkerlab.org/covid_sm_data_bundle. Aggregated statistics derived from the data will be shared on reasonable request to the corresponding author.

REFERENCES

1. Nath A. Long-haul COVID. *Neurology* 2020; 95 (13): 559–60.
2. Gorna R, MacDermott N, Rayner C, *et al.* Long COVID guidelines need to reflect lived experience. *Lancet* 2021; 397 (10273): 455–7.
3. Mendelson M, Nel J, Blumberg L, *et al.* Long-COVID: an evolving problem with an extensive impact. *S Afr Med J* 2020; 111 (1): 10–2.
4. Halpin SJ, Mcivor C, Whyatt G, *et al.* Postdischarge symptoms and rehabilitation needs in survivors of COVID-19 infection: a cross-sectional evaluation. *J Med Virol* 2021; 93 (2): 1013–22.
5. Logue JK, Franko NM, McCulloch DJ, *et al.* Sequelae in adults at 6 months after COVID-19 infection. *JAMA Netw Open* 2021; 4 (2): e210830.
6. Del Rio C, Collins LF, Malani P. Long-term health consequences of COVID-19. *JAMA* 2020; 324 (17): 1723–4.
7. Callard F, Perego E. How and why patients made long COVID. *Soc Sci Med* 2021; 268: 113426.
8. The Lancet. Facing up to long COVID. *Lancet* 2020; 396 (10266): 1861.
9. Marshall M. The four most urgent questions about long COVID. *Nature* 2021; 594 (7862): 168–70.
10. Buonsenso D, Munblit D, De Rose C, *et al.* Preliminary evidence on long COVID in children. *Acta Paediatr* 2021; 110 (7): 2208–11.
11. Fegert JM, Vitiello B, Plener PL, Clemens V. Challenges and burden of the coronavirus 2019 (COVID-19) pandemic for child and adolescent mental

health: a narrative review to highlight clinical and research needs in the acute phase and the long return to normality. *Child Adolesc Psychiatry Ment Health* 2020; 14 (1): 20.

12. Raveendran AV, Jayadevan R, Sashidharan S. Long COVID: an overview. *Diabetes Metab Syndr* 2021; 15 (3): 869–75.

13. Dani M, Dirksen A, Taraborrelli P, *et al.* Autonomic dysfunction in 'long COVID': rationale, physiology and management strategies. *Clin Med (Lond)* 2021; 21 (1): E63–67.

14. Mahase E. COVID-19: what do we know about "long covid"? *BMJ* 2020; 370: m2815.

15. Ladds E, Rushforth A, Wieringa S, *et al.* Persistent symptoms after COVID-19: qualitative study of 114 "long COVID" patients and draft quality principles for services. *BMC Health Serv Res* 2020; 20 (1): 1144.

16. Number of Social Media Users 2025 | Statista. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ Accessed June 9, *2021*.

17. Social Media Use in 2021 | Pew Research Center. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ Accessed June 9, 2021.

18. Shan S, Yan Q, Wei Y. Infectious or recovered? Optimizing the infectious disease detection process for epidemic control and prevention based on social media. *Int J Environ Res Public Health* 2020; 17 (18): 6853–25.

19. Tang L, Bie B, Park SE, Zhi D. Social media and outbreaks of emerging infectious diseases: a systematic review of literature. *Am J Infect Control* 2018; 46 (9): 962–72.

20. Puri N, Coomes EA, Haghbayan H, Gunaratne K. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. *Hum Vaccin Immunother* 2020; 16 (11): 2586–93.

21. Aggarwal NR, Alasnag M, Mamas MA. Social media in the era of COVID-19. *Open Hear* 2020; 7 (2): e001352.

22. Cinelli M, Quattrociocchi W, Galeazzi A, *et al.* The COVID-19 social media infodemic. *Sci Rep* 2020; 10 (1): 16598.

23. Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi MA, Yang Y-C. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. *J Am Med Inform Assoc* 2020; 27 (8): 1310–5.

24. McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A, Srinivasan S. UMLS® knowledge for biomedical language processing. *Bull Med Libr Assoc* 1993; 81 (2): 184–94.

25. covidlonghaulers. https://www.reddit.com/r/covidlonghaulers/ Accessed July 31, 2021.

26. Boe B. PRAW: The Python Reddit API Wrapper. 2017. https://praw.readthedocs.io/en/latest/. Accessed August 24, 2021.

27. Welcome to the NCBO BioPortal | NCBO BioPortal. https://bioportal.bioontology.org/. Accessed April 15, 2021.

28. Mozzicato P. MedDRA: an overview of the medical dictionary for regulatory activities. *Pharm Med* 2009; 23 (2): 65–75.

29. Sarker A. LexExp: a system for automatically expanding concept lexicons for noisy biomedical texts. *Bioinformatics* 2020; doi: 10.1093/bioinformatics/btaa995.

30. Osikomaiya B, Erinoso O, Wright KO, *et al.* 'Long COVID': persistent COVID-19 symptoms in survivors managed in Lagos State, Nigeria. *BMC Infect Dis* 2021; 21 (1): 1–7.

31. Huang C, Huang L, Wang Y, *et al.* 6-Month consequences of COVID-19 in patients discharged from hospital: a cohort study. *Lancet* 2021; 397 (10270): 220–32.

32. Sykes DL, Holdsworth L, Jawad N, Gunasekera P, Morice AH, Crooks MG. Post-COVID-19 symptom burden: what is long-COVID and how should we manage it? *Lung* 2021; 199 (2): 113–9.

33. Mandal S, Barnett J, Brill SE, *et al.*; ARC Study Group. 'Long-COVID': a cross-sectional study of persisting symptoms, biomarker and imaging abnormalities following hospitalisation for COVID-19. *Thorax* 2021; 76 (4): 396–8.

34. Carfì A, Bernabei R, Landi F; for the Gemelli Against COVID-19 Post-Acute Care Study Group. Persistent symptoms in patients after acute COVID-19. *JAMA* 2020; 324 (6): 603–5.

35. Orrù G, Bertelloni D, Diolaiuti F, *et al.* Long-COVID syndrome? A study on the persistence of neurological, psychological and physiological symptoms. *Healthcare* 2021; 9 (5): 575.

36. Garrigues E, Janvier P, Kherabi Y, *et al.* Post-discharge persistent symptoms and health-related quality of life after hospitalization for COVID-19. *J Infect* 2020; 81 (6):e4–6.

37. Lopez-Leon S, Wegman-Ostrosky T, Perelman C, *et al.* More than 50 long-term effects of COVID-19: a systematic review and meta-analysis. *Sci Rep* 2021; 11 (1): 16144.