

Trend and co-occurrence network study of symptoms through social media: an example of COVID-19

Jiageng Wu^{1,2*}, Lumin Wang^{1*}, Yining Hua^{3,4}, Minghui Li^{1,2}, Li Zhou^{3,4}, David W Bates⁴,
Jie Yang^{1,2#}

¹ Department of Big Data in Health Science School of Public Health, and Center of Clinical Big Data and Analytics of The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China.

² The Key Laboratory of Intelligent Preventive Medicine of Zhejiang Province, Hangzhou, Zhejiang, China.

³ Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

⁴ Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA, USA

* *Equally contribution*

Corresponding author

Corresponding author:

Jie Yang, Ph.D.

Department of Big Data in Health Science School of Public Health, and Center of Clinical Big Data and Analytics of The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China

Address: 866 Yuhangtang Rd, Hangzhou, P.R. China 310058

Email: y@zju.edu.cn, jielynlp@gmail.com

Telephone: +86-19157731185

Manuscript word count: 2968 words

Key points

Questions: What are the epidemic characteristics and relationships of COVID-19 symptoms that have been extensively reported on social media?

Findings: This retrospective cohort study of 948,478 related tweets (February 2020 to April 2022) from 689,551 users identified 201 self-reported COVID-19 symptoms from 10 affected systems, mitigating the potential missing information in hospital-based epidemiologic studies due to many patients not being timely diagnosed and treated. Coma, anosmia, taste sense altered, and dyspnea were less common in participants infected during Omicron prevalence than in Delta. Symptoms that affect the same system have high co-occurrence. Frequent co-occurrences occurred between symptoms and systems corresponding to specific disease progressions, such as palpitations and dyspnea, alopecia and impotence.

Meaning: Trend and network analysis in social media can mine dynamic epidemic characteristics and relationships between symptoms in emergent pandemics.

Abstract

Importance: COVID-19 is a multi-organ disease with broad-spectrum manifestations. Clinical data-driven research can be difficult because many patients do not receive prompt diagnoses, treatment, and follow-up studies. Social media's accessibility, promptness, and rich information provide an opportunity for large-scale and long-term analyses, enabling a comprehensive symptom investigation to complement clinical studies.

Objective: Present an efficient workflow to identify and study the characteristics and co-occurrences of COVID-19 symptoms using social media.

Design, Setting, and Participants: This retrospective cohort study analyzed 471,553,966 COVID-19-related tweets from February 1, 2020, to April 30, 2022. A comprehensive lexicon of symptoms was used to filter tweets through rule-based methods. 948,478 tweets with self-reported symptoms from 689,551 Twitter users were identified for analysis.

Main Outcomes and Measures: The overall trends of COVID-19 symptoms reported on Twitter were analyzed (separately by the Delta strain and the Omicron strain) using weekly new numbers, overall frequency, and temporal distribution of reported symptoms. A co-occurrence network was developed to investigate relationships between symptoms and affected organ systems.

Results: The weekly quantity of self-reported symptoms has a high consistency (0.8528, $P < 0.0001$) and one-week leading trend (0.8802, $P < 0.0001$) with new infections in four countries.

We grouped 201 common symptoms (mentioned ≥ 10 times) into 10 affected systems. The frequency of symptoms showed dynamic changes as the pandemic progressed, from typical respiratory symptoms in the early stage to more musculoskeletal and nervous symptoms at later

stages. When comparing symptoms reported during the Delta strain versus the Omicron variant, significant changes were observed, with dropped odd ratios of coma (95%CI 0.55-0.49, $P < 0.01$) and anosmia (95%CI, 0.6-0.56), and more pain in the throat (95%CI, 1.86-1.96) and concentration problems (95%CI, 1.58-1.70). The co-occurrence network characterizes relationships among symptoms and affected systems, both intra-systemic, such as cough and sneezing (respiratory), and inter-systemic, such as alopecia (integumentary) and impotence (reproductive).

Conclusions and Relevance: We found dynamic COVID-19 symptom evolution through self-reporting on social media and identified 201 symptoms from 10 affected systems. This demonstrates that social media's prevalence trends and co-occurrence networks can efficiently identify and study public health problems, such as common symptoms during pandemics.

Introduction

The global coronavirus disease 2019 pandemic (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has resulted in more than 616 million infections and 6.54 million deaths as of 28 September 2022.¹ Furthermore, the pandemic is still ongoing, and its catastrophic impact may continue to grow and last for years. To broaden the understanding of this disease, relevant studies have been increasingly emerging, from determining molecular structures^{2,3} to developing drugs and vaccines.⁴⁻⁶ Concurrently, clinicians have endeavored to analyze clinical symptoms to guide therapeutic strategies.⁷ Public health officials have also tried to investigate the prevalence of symptoms to utilize the findings to provide precise prevention and control strategies for both people and governments.^{8,9}

As a popular communication tool and public discussion platform, social media such as Twitter has permeated every aspect of our daily lives. Especially during the pandemic, social media played an essential role in information generation, dissemination, and consumption.^{10,11} There has been emerging COVID-19-related research based on social media. Such studies include topics in infodemics, public attitudes, detection or prediction of confirmed cases, and government responses to the pandemic¹². However, they mainly focused on thematic analysis^{13,14} or sentiment analysis.^{15,16} Only a few studies analyzed the symptoms and their epidemic-related characteristics.¹⁷⁻¹⁹ Moreover, these studies mainly conducted distribution and trend analyses in the early months of the pandemic rather than long-term, comprehensive investigations.

Current understandings of COVID-19 symptoms are primarily established on clinical data from medical institutions²⁰⁻²², such as electronic health records (EHRs). However, nearly 80% of

patients with asymptomatic or mild symptoms are not promptly or never clinically diagnosed and treated²³⁻²⁵, leading to potential missing information for mild and early symptoms. In addition, privacy policies on patient data have slowed cross-institutional cooperation and thorough studies of the pandemic on a large scale.²⁶ Due to limited data sizes and sample diversity, current COVID-19 symptom network analyses only include a few typical symptoms, and do not construct a holistic network of comprehensive symptoms and affected systems.^{27,28}

To attempt to address this research gap, we propose an efficient workflow for tracking and analyzing the general prevalence status and relationships of COVID-19 symptoms using social media.

Methods

Data collection

We selected non-retweeted English tweets related to COVID-19 using unique tweet identifiers from a widely used open-source COVID-19 tweet database.^{29,30} The tweets were identified by Twitter's trending topics and selected keywords associated with COVID-19, such as *COVID-19* and *SARS-COV-2*. We downloaded 471,553,966 related tweets across 27 months, from February 1, 2020, to April 30, 2022, using Twitter's Application Programming Interface (API).

Symptom lexicon

Based on current literature, we built a comprehensive and hierarchical COVID-19 symptoms lexicon containing synonyms of symptoms and their affected body parts.³¹⁻³⁵ Specifically, we appended colloquial variants frequently found on social media (**eMethod 1**). In addition, we

grouped symptoms according to the affected organs and systems into 10 families^{36,37}: cardiovascular, digestive, integumentary, musculoskeletal, nervous, reproductive, respiratory, urinary, sensory, and systemic. The final symptom lexicon contains 10 affected organs/systems, 257 symptoms, and 1808 synonyms (**Supplement Files 2**).

Text preprocessing and rule-based filtering

To identify tweets with self-reported symptoms for subsequent analysis, we designed a three-step method that can be roughly summarized into filtering tweets with strict COVID-19 keywords, text cleaning, and matching of self-reported symptoms (**eMethod 2**). The overall workflow of this study is shown in **Figure 1**.

Trend analysis on the quantity of new COVID-19-related tweets

We compared weekly numbers of new COVID-19 tweets to new cases in countries with the most Twitter users. A survey on Statista shows that as of Jan 2021³⁸, the top 4 countries with the most Twitter users and use English as their primary language are the United States (US), the United Kingdom (UK), the Philippines, and Canada (**eTable 2**). We used new COVID-19 cases in these countries reported by the World Health Organization (WHO) to be a rough representation of COVID-19 new cases (**Supplement Files 3**). We calculated weekly numbers of new tweets for both before and after the filtering. We also computed their Pearson correlation coefficient (P_{cc}) with the number of new cases to examine whether there was a statistically significant association between COVID-19 severity and public response.

Overall distribution and dynamic frequency analysis of symptoms

Based on the COVID-19 symptom lexicon, we counted occurrences of each symptom by

matching their synonyms against the filtered tweets datasets. Multiple mentions of the same symptom in one tweet were counted as one. To explore dynamic changes in symptom distribution with time, we calculated each symptom's weekly frequency, normalized by the number of all self-reporting tweets. We also calculated the normalized frequency for each affected system.

Comparison of symptoms prevalence status between different strains

COVID-19 has several variants that present different epidemic characteristics³⁹, such as the highly transmissible B.1.617.2 (Delta) variant^{40,41} and B.1.1.529 (Omicron) variant⁴², which have led to rapid global rises. In this section, we compare self-reported symptom frequencies between Delta and Omicron. We extracted tweets from June 1, 2021, to Nov 27, 2021, when Delta was the globally dominant variant^{36,43,44} to represent Delta. Respectively, we extracted tweets from Dec 20, 2021, to Apr 30, 2022³⁶ to represent Omicron.

We extracted symptoms from the two groups of tweets and selected those with $\geq 1\%$ frequency as common symptoms. Then, we used the Chi-square test to calculate odds ratios (ORs) for Delta versus Omicron to assess the approximate prevalence differences of these common symptoms in two periods. Since a patient can get Delta in the Omicron-dominated period, this method calculates the odds of detecting a symptom among infected participants during the Delta-dominated period compared to the Omicron period.

Network analysis

A COVID-19 patient may have multiple symptoms and report them simultaneously. Based on the symptom lexicon, we matched each symptom against each tweet to create a dataset $\mathbf{X} = [x_1, x_2, \dots, x_n] \in R^{n \times m}$, where $x_i = [d_{i1}, d_{i2}, \dots, d_{im}]$. d_{ij} is a binary feature that represents whether

tweet x_i mentions symptom j ; m and n represent the numbers of symptoms and tweets, respectively.

To quantitatively explore the strength of co-occurrence between two symptoms, we built symptom vectors \mathbf{V} , where $\mathbf{V} = \mathbf{X}^T = [v_1, v_2, \dots, v_m] \in R^{m \times n}$, meaning that each dimension of v_x is a binary feature that indicates whether the symptom x was mentioned in tweet i . The co-occurrence strength is modeled by the similarity between the two symptom vectors, for which we adopted cosine similarity as the metric. In conclusion, the co-occurrence C between v_x and v_y can be modeled by the following equation:

$$C(v_x, v_y) = \frac{\sum_{i=1}^n v_{xi} v_{yi}}{\sqrt{\sum_{i=1}^n v_{xi}^2} \sqrt{\sum_{i=1}^n v_{yi}^2}}$$

Based on the model, we constructed a weighted co-occurrence network of COVID-19 symptoms, where nodes represent symptoms and edges capture the co-occurrence strength between symptom pairs. We used Gephi⁴⁵ and ForceAtlas2 algorithm⁴⁶ to visualize the symptom network.

Results

We selected 948,478 unique COVID-19-related tweets with self-reported symptoms to conduct these studies.

Weekly trends of self-reporting tweets

We observed that weekly changes in self-reporting tweets were roughly consistent with the trends of new cases in the four selected countries (**Figure 2A**). The P_{cc} between the two trends is 0.8528 ($P < 0.0001$), higher than the P_{cc} between new cases and the unfiltered COVID-19-related tweets (0.3235, $P = 0.0004$, **eFigure 1**). Moreover, self-reporting tweets showed a significant leading trend compared to the new cases when the leading time was set to one week. Such a trend had a higher correlation ($P_{cc} = 0.8802$, $P < 0.0001$) than when no time difference was set.

Distribution of COVID-19 symptoms and affected organs/systems

In all, 245 symptoms were mentioned a total of 1197,733 times in 948,478 tweets. A total of 201 symptoms from 10 affected systems were mentioned in ≥ 10 tweets. The distribution of different systems and their related symptoms are hierarchically visualized in **eFigure 2**. Notably, systemic symptoms accounted for 42% of the total number of symptom occurrences, followed by respiratory (33%), digestive (7%), sensory (6%), musculoskeletal (4%), nervous (4%), integumentary (2%), cardiovascular (1%), reproductive (0.202%) and urinary (0.0645%) symptoms.

Frequency of the common COVID-19 symptoms and affected systems

Overall, 20 common symptoms have more than a 1% frequency (**Table 1**) (more details in **Supplement Files 4**). Note that the WHO report was based on 55,924 laboratory-confirmed cases from China in the early stage of COVID-19.⁴⁷ The data of Delta and Omicron were extracted and calculated from our dataset in the corresponding period.

Figure 2B and **Figure 2C** show the weekly frequency of COVID symptoms and affected systems. The frequency of symptoms shows dynamic changes with the progression of the pandemic and has some distinct waves, respectively. In the early stage of COVID-19, cough, fever, and sneezing were the major symptoms, while other symptoms were rarely reported. With the progression of the pandemic, more symptoms, such as taste sense altered, chill, and anosmia, started to emerge. Respiratory symptoms were most common initially, accounting for more than 80% at one time, then gradually decreasing to about 40%. In contrast, the frequency of systemic, musculoskeletal, and nervous symptom mentions showed increasing trends. Frequencies of

different symptoms gradually stabilized, with fluctuations associated with hotspot issues and the emergence of new variants.

Distribution difference in symptoms between COVID-19 variants

The 209,074 tweets from June 1, 2021 to November 27, 2021 were placed in the Delta group, while 244,960 tweets from December 20, 2020 to April 30, 2021 were selected for the Omicron group. **Table 1** shows their top common symptoms and corresponding frequencies. **Figure 3** shows the ORs of common symptoms for Delta versus Omicron.

The top 20 symptoms of Omicron and Delta were roughly the same, but nasal congestion replaced coma as one of the top 20 symptoms of Omicron. Among these 21 symptoms, 8 were significantly ($P < 0.01$) less prevalent (all $P < 0.01$) among individuals infected during the Omicron period than Delta (top-5 OR: coma 0.52 [0.55-0.49], anosmia 0.58 [0.6-0.56], taste sense altered 0.66 [0.68-0.64], dyspnea 0.83 [0.85-0.81], chill 0.86 [0.89-0.82]), and 10 were significantly more likely to occur in Omicron patients than Delta (top-5 OR: pain in throat 1.91 [1.86-1.96], concentration problems 1.64 [1.58-1.70], nasal congestion 1.47 [1.38-1.55], rhinorrhea 1.37 [1.33-1.41], cough 1.21 [1.19-1.23])(More details in **eTable 3**).

The co-occurrence network of COVID-19 symptoms

To simplify the co-occurrence network, we selected the top 100 symptoms by their overall distribution. The final network has 100 nodes with 2654 edges (**Figure 4**). Overall, the symptoms in this network show a clustering tendency according to the affected system, and the common symptoms are roughly distributed in the central region. Though systemic and musculoskeletal symptoms were not the leading part of the network, they are mainly in the center of the network

and linked to the symptoms of different systems. Some outliers fall out of the clustering region of their theoretically affected systems. For example, palpitations, a cardiovascular symptom, locates at the center of the network next to systemic and musculoskeletal symptoms. Impotence, the only reproductive symptom with a high occurrence rate, and nocturnal enuresis, the only urinary symptom, located at the network border, demonstrating that co-occurrence with other symptoms were relatively low. Both intra- and inter-systemic symptoms had strong co-occurrences, such as chills and fever (both systemic symptoms), palpitations (cardiovascular) and dyspnea (respiratory), etc. For the readers to further explore the co-occurrences of a specific symptom, we provide an online visualization of the network (<https://jgwu.top/COVID19-Symptoms-Twitter/network/>).

Discussion

In this work, we presented a novel workflow to investigate the symptom characteristics of an emergent pandemic using social media. We curated a hierarchical symptom lexicon that handles social media colloquialism and maps symptoms to their affected systems. This lexicon can be used in further social media-based medical research. We have also contributed a comprehensive co-occurrence network for COVID-19 symptoms for further exploration. To the best of our knowledge, this is the first dynamic prevalence status and network analysis of COVID-19 symptoms using large-scale and long-term social media data. This workflow can aid clinical professionals in monitoring unusual co-occurrent symptom patterns to promote pathogenesis studies. It is also promising in studying other emergent epidemics, given the accessibility and timeliness of social media.

Through the time trend analysis, we observed consistency between the trend of self-reporting

tweets and new COVID cases ($P_{cc}=0.8054$). This suggests a highly positive significant correlation between the severity of the pandemic and the number of self-reported symptoms on social media. Masri et al. found that new case trends could be predicted one week ahead based on related tweets for the 2015 Zika epidemic.⁴⁸ In correspondence and beyond, we found a highly correlated one-week leading trend of symptom-related tweets compared to new cases ($P_{cc}=0.8802, p<0.0001$) for COVID-19. This further demonstrates the sensitivity of social media and emphasizes the effectiveness of studying symptoms using social media in timely monitoring and prediction of pandemic status. Meanwhile, small fluctuations in the trends reflect public concerns with hotspot issues such as government policies and measures regarding the pandemic. For example, **Figure 2A** shows that the presidential election and Trump testing positive triggered increases in self-reporting tweets. This could be attributed to people discussing relevant problems and bringing up their own experiences, including symptoms. The insights gained from this type of trend analysis could help officials better guide and warn the public during pandemics. Readers can refer to our previous study for a more detailed investigation of the influence of hotspot issues on symptom reports.⁴⁹

The common symptoms and their occurrence/prevalence ranks identified in our study are mostly in accordance with WHO reports but with different frequencies. These differences can be partially attributed to the different granularity and definition of symptoms. For example, cough in the WHO report only refers to dry cough, whereas wet cough is often correlated with sputum production.⁴⁷ Such strict definitions are less suitable for self-reported social media data than traditional clinical studies. Using the adapted symptom lexicon, we identified a few symptoms that

were not taken seriously in the WHO's early reports, such as taste sense altered, anosmia, and nausea⁵⁰⁻⁵². We also noticed some relatively infrequent symptoms, such as alopecia (occurrence: 5373) and impotence (occurrence: 2027). Recent studies have confirmed that SARS-CoV-2 may affect the expression of androgen and corrupt the physiological pathways involved in regulating erection.⁵³⁻⁵⁵ Having learned from the UK government's experience of being urged by general practitioners to update the official COVID-19 symptom list to eliminate confusion^{56,57}, policymakers should be aware that timely updates on the disease are essential to reassure the public, control the disease and better manage patients with specific complications.

Figure 2 shows that symptom prevalence varied over time along with the virus evolution. As the key receptors of SARS-CoV-2 are highly co-expressive in the respiratory tract⁵⁸⁻⁶⁰, the initial symptoms are mainly respiratory and systemic symptoms caused by inflammation. However, over time, extensive self-reports of multiple symptoms from different systems confirmed that COVID-19 is a multi-organ disease⁶¹. At the later stage of the pandemic, there are increasing reports of persistent symptoms after COVID-19, such as fatigue, concentration problems, and limb pain (muscle/joint).^{62,63} Notably, consistent with recent findings on the increased risks of cardiovascular diseases⁶⁴ and long neuropsychiatric symptoms⁶⁵, our results show a burst of attention to nervous and cardiovascular symptoms on social media in January 2022, which have continued growing. This alerts us to the emerging prolonged signs (long-COVID)⁶⁶ and their chronic burden on the nervous and cardiovascular systems.

Through the Chi-square test and **Figure 3**, we found that compared to Delta, as reported by the general population, Omicron has (1) lower ORs of severe symptoms, such as coma and dyspnea;

(2) higher ORs for flu-like symptoms, such as pain in the throat, concentration problems, nasal congestion, and rhinorrhea; and (3) lower ORs of some typical COVID symptoms, such as anosmia and taste sense altered.^{36,67,68} This finding confirms that the Omicron is much more transmissible than previous variants but has less severe symptoms.^{69,70}

The network of COVID symptoms and affected systems, built on massive data and a comprehensive lexicon, contains more extensive information than previous studies.^{27,28} While symptoms of the same system have higher co-occurrences, we did observe inter-system co-occurrences consistent with clinical studies. For example, coma exhibits strong relationships with respiratory symptoms in our networks, especially dyspnea, because the hypoxic/metabolic changes caused by intense inflammatory response trigger cytokine storm and may further result in coma and encephalopathy.⁷¹ We also found unusual co-occurrences. For example, palpitations as a cardiovascular symptom strongly correlate with dyspnea and dizziness (respiratory and systemic).⁷² Impotence, a reproductive symptom, has the strongest correlation with alopecia (an integumentary symptom), likely due to the SARS-CoV-2 invasion on androgen expression. They are also the typical long-COVID symptom among non-hospitalized patients.⁷³ These strong relationships among unexpected group symptoms may point to new foci of disease progression or alert the potential risk of co-occurrent symptoms.

Limitations

We acknowledge that our study has limitations. First, although we have reviewed substantial studies to construct a lexicon that is as comprehensive as possible, it inevitably misses some colloquial variants of the symptoms due to the noisy nature of Twitter. Second, the self-reported

symptoms and cases are not laboratory-confirmed results. Moreover, some of our analyses could be biased. For example, we would expect more Delta patients in real life than in our study since people could still get Delta in the Omicron-dominated period. Therefore, we explicitly point out that our comparison is an estimation. Third, like every other public health study based on social media data, our study has potential cohort bias as the demographic distribution of social media does not represent that of the whole population.

Conclusions

We developed a novel workflow to explore the dynamic characteristics of pandemic symptoms through social media. Using symptom analysis, we performed a large-scale and long-term social media-based study on COVID-19 and identified 201 symptoms from 10 systems. Compared to clinical data-based studies, we found a different symptom prevalence reported by a population of predominantly mild symptom patients. Evaluations like this can complement clinical studies to depict a more holistic picture of COVID-19 symptoms. The network reveals unusual co-occurrent symptom patterns, which may enable downstream pathogenesis studies. Thanks to the accessibility and timeliness of social media, this workflow is also promising in contributing to future public health studies, such as studying other emergent epidemics.

Data availability.

The code of this study is available at: <https://github.com/Dragon-Wu/COVID19-Symptoms-Twitter>.

Acknowledgments

This study has no funding support.

Author Contributions

J.W, L.W, and M.L performed data analysis and drafted the manuscript. J.Y. designed the study.

J.W, L.W, and M.L developed the symptom lexicon. Y.H prepared the data, helped draft and revised the manuscript. L.Z., J.Y., and D.W.B. provided critical reviews. All authors reviewed the manuscript. J.W. takes responsibility for the integrity of the work.

Conflict of Interest Disclosures

Dr. Bates reports grants and personal fees from EarlySense, personal fees from CDI Negev, equity from ValeraHealth, equity from Clew, equity from MDClone, personal fees and equity from AESOP, personal fees and equity from FeelBetter, and grants from IBM Watson Health, outside the submitted work.

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*. 2020;20(5):533-534. doi:10.1016/s1473-3099(20)30120-1
2. Yao HP, Song YT, Chen Y, et al. Molecular Architecture of the SARS-CoV-2 Virus. *Cell*. Oct 29 2020;183(3):730-+.
3. Wrapp D, Wang NS, Corbett KS, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*. Mar 13 2020;367(6483):1260-+.
4. Yin WC, Mao CY, Luan XD, et al. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science*. Jun 26 2020;368(6498):1499-+.

5. Zhou YD, Wang F, Tang J, Nussinov R, Cheng FX. Artificial intelligence in COVID-19 drug repurposing. *Lancet Digit Health*. Dec 2020;2(12):E667-E676.
6. Bernal JL, Andrews N, Gower C, et al. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *New Engl J Med*. Aug 12 2021;385(7):585-594.
7. Crook H, Raza S, Nowell J, Young M, Edison P. Long covid-mechanisms, risk factors, and management. *Bmj-Brit Med J*. Jul 26 2021;374
8. Budd J, Miller BS, Manning EM, et al. Digital technologies in the public-health response to COVID-19. *Nat Med*. Aug 2020;26(8):1183-1192. doi:10.1038/s41591-020-1011-4
9. Escandon K, Rasmussen AL, Bogoch, II, et al. COVID-19 false dichotomies and a comprehensive review of the evidence regarding public health, COVID-19 symptomatology, SARS-CoV-2 transmission, mask wearing, and reinfection. *BMC Infect Dis*. Jul 27 2021;21(1):710. doi:10.1186/s12879-021-06357-4
10. Aiello AE, Renson A, Zivich PN. Social Media- and Internet-Based Disease Surveillance for Public Health. *Annu Rev Publ Health*. 2020;41:101-118.
11. Al-Surimi K, Khalifa M, Bahkali S, EL-Metwally A, Househ M. The Potential of Social Media and Internet-Based Data in Preventing and Fighting Infectious Diseases: From Internet to Twitter. *Adv Exp Med Biol*. 2017;972:131-139.
12. Tsao S-F, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA. What social media told us in the time of COVID-19: a scoping review. *The Lancet Digital Health*. 2021;3(3):e175-e194. doi:10.1016/s2589-7500(20)30315-0
13. Boon-Itt S, Skunkan Y. Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. *Jmir Public Health and Surveillance*. Oct-Dec 2020;6(4):245-261.
14. Li LY, Zhou JY, Ma ZH, Bensi MT, Hall MA, Baecher GB. Dynamic assessment of the COVID-19 vaccine acceptance leveraging social media data. *Journal of Biomedical Informatics*. May 2022;129
15. Xue J, Chen JX, Hu R, et al. Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *Journal of Medical Internet Research*. Nov 25 2020;22(11)
16. Hussain A, Tahir A, Hussain Z, et al. Artificial Intelligence-Enabled Analysis of Public Attitudes on Facebook and Twitter Toward COVID-19 Vaccines in the United Kingdom and the United States: Observational Study. *Journal of Medical Internet Research*. Apr 5 2021;23(4)
17. Alanazi E, Alashaikh A, Alqurashi S, Alanazi A. Identifying and Ranking Common COVID-19 Symptoms From Tweets in Arabic: Content Analysis. *Journal of Medical Internet Research*. Nov 18 2020;22(11)
18. Sarabadani S, Baruah G, Fossat Y, Jeon J. Longitudinal Changes of COVID-19 Symptoms in Social Media: Observational Study. *Journal of Medical Internet Research*. Feb 16 2022;24(2)
19. Guo JW, Radloff CL, Wawrzynski SE, Cloyes KG. Mining twitter to explore the emergence of COVID-19 symptoms. *Public Health Nurs*. Nov 2020;37(6):934-940.
20. Grant MC, Geoghegan L, Arbyn M, et al. The prevalence of symptoms in 24,410 adults infected by the novel coronavirus (SARS-CoV-2; COVID-19): A systematic review and meta-analysis of 148 studies from 9 countries. *Plos One*. Jun 23 2020;15(6)
21. Amin MT, Hasan M, Bhuiya NMMA. Prevalence of Covid-19 Associated Symptoms, Their Onset and Duration, and Variations Among Different Groups of Patients in Bangladesh. *Frontiers in Public Health*. Sep 29 2021;9
22. Wang LQ, Foer D, MacPhaul E, Lo YC, Bates DW, Zhou L. PASClex: A comprehensive post-acute sequelae of COVID-19 (PASC) symptom lexicon derived from electronic health record clinical notes. *Journal of Biomedical Informatics*. Jan 2022;125
23. Ma Q, Liu J, Liu Q, et al. Global percentage of asymptomatic SARS-CoV-2 infections among the tested

population and individuals with confirmed COVID-19 diagnosis: a systematic review and meta-analysis. 2021;4(12):e2137257-e2137257.

24. Sah P, Fitzpatrick MC, Zimmer CF, et al. Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis. 2021;118(34):e2109229118.

25. Wu Z, McGoogan JMJ. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. 2020;323(13):1239-1242.

26. Dagliati A, Malovini A, Tibollo V, Bellazzi R. Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview. *Brief Bioinform*. Mar 22 2021;22(2):812-822. doi:10.1093/bib/bbaa418

27. Millar JE, Neyton L, Seth S, et al. Distinct clinical symptom patterns in patients hospitalised with COVID-19 in an analysis of 59,011 patients in the ISARIC-4C study. *Scientific Reports*. 2022/04/27 2022;12(1):6843. doi:10.1038/s41598-022-08032-3

28. Fernández-de-las-Peñas C, Varol U, Gómez-Mayordomo V, Cuadrado ML, Valera-Calero JA. The relevance of headache as an onset symptom in COVID-19: a network analysis of data from the LONG-COVID-EXP-CM multicentre study. *Acta Neurologica Belgica*. 2022/08/01 2022;122(4):1093-1095. doi:10.1007/s13760-022-01998-x

29. Chen E, Lerman K, Ferrara E. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance*. 2020;6(2):e19273.

30. Lopez CE, Gallemore C. An augmented multilingual Twitter dataset for studying the COVID-19 infodemic. *Social Network Analysis and Mining*. 2021;11(1):1-14.

31. Wang L, Blackley SV, Blumenthal KG, et al. A dynamic reaction picklist for improving allergy reaction documentation in the electronic health record. *J Am Med Inform Assoc*. Jun 1 2020;27(6):917-923. doi:10.1093/jamia/ocaa042

32. Goss FR, Lai KH, Topaz M, et al. A value set for documenting adverse reactions in electronic health records. *J Am Med Inform Assoc*. Jun 1 2018;25(6):661-669. doi:10.1093/jamia/ocx139

33. Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi MA, Yang YC. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. *J Am Med Inform Assoc*. Aug 1 2020;27(8):1310-1315. doi:10.1093/jamia/ocaa116

34. 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*. Aug 9 2021;11(1):16144. doi:10.1038/s41598-021-95565-8

35. Mao L, Jin H, Wang M, et al. Neurologic Manifestations of Hospitalized Patients With Coronavirus Disease 2019 in Wuhan, China. *JAMA Neurol*. Jun 1 2020;77(6):683-690. doi:10.1001/jamaneurol.2020.1127

36. Menni C, Valdes AM, Polidori L, et al. Symptom prevalence, duration, and risk of hospital admission in individuals infected with SARS-CoV-2 during periods of omicron and delta variant dominance: a prospective observational study from the ZOE COVID Study. *Lancet*. Apr 23 2022;399(10335):1618-1624. doi:10.1016/S0140-6736(22)00327-0

37. SEER-Training(NIH). Review: Introduction to the Human Body. <https://training.seer.cancer.gov/anatomy/body/review.html#:~:text=A%20system%20is%20an%20organization,urinary%2C%20and%20the%20reproductive%20system>.

38. statista. Leading countries based on number of Twitter users as of January 2021. <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>

39. WHO. Tracking SARS-CoV-2 variants. <https://www.who.int/activities/tracking-SARS-CoV-2-variants>

40. Mlcochova P, Kemp SA, Dhar MS, et al. SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion.

Nature. Nov 4 2021;599(7883):114-+. doi:10.1038/s41586-021-03944-y

41. del Rio C, Malani PN, Omer SB. Confronting the Delta Variant of SARS-CoV-2, Summer 2021. *Jama-J Am Med Assoc*. Sep 21 2021;326(11):1001-1002. doi:10.1001/jama.2021.14811
42. Hu J, Peng P, Cao X, et al. Increased immune escape of the new SARS-CoV-2 variant of concern Omicron. *Cell Mol Immunol*. Feb 2022;19(2):293-295. doi:10.1038/s41423-021-00836-z
43. US C-t. CDC Museum COVID-19 Timeline. <https://www.cdc.gov/museum/timeline/covid19.html>
44. Torjesen I. Covid-19: Delta variant is now UK's most dominant strain and spreading through schools. British Medical Journal Publishing Group; 2021.
45. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. 2009:361-362.
46. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One*. 2014;9(6):e98679. doi:10.1371/journal.pone.0098679
47. WHO. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19) <https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf>
48. Masri S, Jia JF, Li C, et al. Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic. *BMC public health*. Jun 14 2019;19doi:ARTN 761 10.1186/s12889-019-7103-8
49. Hua Y, Jiang H, Lin S, et al. Using Twitter data to understand public perceptions of approved versus off-label use for COVID-19-related medications. *Journal of the American Medical Informatics Association*. 2022;doi:10.1093/jamia/ocac114
50. Menni C, Valdes AM, Freidin MB, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat Med*. Jul 2020;26(7):1037-+. doi:10.1038/s41591-020-0916-2
51. Benezit F, Le Turnier P, Declerck C, et al. Utility of hyposmia and hypogeusia for the diagnosis of COVID-19. *Lancet Infect Dis*. Sep 2020;20(9):1014-1015. doi:10.1016/S1473-3099(20)30297-8
52. Andrews PLR, Cai WG, Rudd JA, Sanger GJ. COVID-19, nausea, and vomiting. *J Gastroen Hepatol*. Mar 2021;36(3):646-656. doi:10.1111/jgh.15261
53. Kraljevic Z, Searle T, Shek A, et al. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artificial intelligence in medicine*. Jul 2021;117:102083. doi:10.1016/j.artmed.2021.102083
54. Katz J, Yue S, Xue W, Gao H. Increased odds ratio for erectile dysfunction in COVID-19 patients. *J Endocrinol Invest*. Apr 2022;45(4):859-864. doi:10.1007/s40618-021-01717-y
55. Kaynar M, Gomes ALQ, Sokolakis I, Gul M. Tip of the iceberg: erectile dysfunction and COVID-19. *Int J Impot Res*. Mar 2022;34(2):152-157. doi:10.1038/s41443-022-00540-0
56. Iacobucci G. Covid-19: UK adds sore throat, headache, fatigue, and six other symptoms to official list. *Bmj-Brit Med J*. Apr 4 2022;377
57. Mahase E. Covid-19: GPs urge government to clear up confusion over symptoms. *Bmj-Brit Med J*. Jun 28 2021;373
58. Yan RH, Zhang YY, Li YN, Xia L, Guo YY, Zhou Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science*. Mar 27 2020;367(6485):1444-+.
59. Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*. Apr 16 2020;181(2):271-+.

60. Sungnak W, Huang N, Becavin C, et al. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat Med*. May 2020;26(5):681-+.
61. Kumar A, Narayan RK, Prasoon P, et al. COVID-19 Mechanisms in the Human Body-What We Know So Far. *Front Immunol*. Nov 1 2021;12
62. Blomberg B, Mohn KGI, Brokstad KA, et al. Long COVID in a prospective cohort of home-isolated patients. *Nat Med*. Sep 2021;27(9):1607-+.
63. Blomberg B, Cox RJ, Langeland N. Long COVID: A growing problem in need of intervention. *Cell Rep Med*. Mar 15 2022;3(3)
64. Xie Y, Xu E, Bowe B, Al-Aly Z. Long-term cardiovascular outcomes of COVID-19. *Nat Med*. Mar 2022;28(3):583-590. doi:10.1038/s41591-022-01689-3
65. Boldrini M, Canoll PD, Klein RS. How COVID-19 Affects the Brain. *Jama Psychiat*. Jun 2021;78(6):682-683. doi:10.1001/jamapsychiatry.2021.0500
66. Mehandru S, Merad M. Pathological sequelae of long-haul COVID. *Nat Immunol*. Feb 2022;23(2):194-202.
67. Molteni E, Sudre CH, Canas LDS, et al. Illness Characteristics of COVID-19 in Children Infected with the SARS-CoV-2 Delta Variant. *Children (Basel)*. May 3 2022;9(5)doi:10.3390/children9050652
68. Menni C, Valdes AM, Polidori L, et al. Symptom prevalence, duration, and risk of hospital admission in individuals infected with SARS-CoV-2 during periods of omicron and delta variant dominance: a prospective observational study from the ZOE COVID Study. *Lancet*. Apr 23 2022;399(10335):1618-1624. doi:10.1016/S0140-6736(22)00327-0
69. Mannar D, Saville JW, Zhu X, et al. SARS-CoV-2 Omicron variant: Antibody evasion and cryo-EM structure of spike protein-ACE2 complex. *Science*. Feb 18 2022;375(6582):760-+. doi:10.1126/science.abn7760
70. Hui KPY, Ho JCW, Cheung MC, et al. SARS-CoV-2 Omicron variant replication in human bronchus and lung ex vivo. *Nature*. Mar 24 2022;603(7902):715-+. doi:10.1038/s41586-022-04479-6
71. Garg RK, Paliwal VK, Gupta A. Encephalopathy in patients with COVID-19: A review. *J Med Virol*. Jan 2021;93(1):206-222. doi:10.1002/jmv.26207
72. Bisaccia G, Ricci F, Recce V, et al. Post-Acute Sequelae of COVID-19 and Cardiovascular Autonomic Dysfunction: What Do We Know? *J Cardiovasc Dev Dis*. Nov 15 2021;8(11)doi:10.3390/jcdd8110156
73. Wise J. Long covid: Hair loss and sexual dysfunction are among wider symptoms, study finds. *BMJ*. Jul 27 2022;378:o1887. doi:10.1136/bmj.o1887

Figure Legend

- Figure 1. The overall workflow
- Figure 2. Weekly numbers of self-reporting tweets and weekly trends of the frequency of symptoms and affected systems
- Figure 3. Distribution difference in common symptoms between Delta and Omicron
- Figure 4. The co-occurrence network of different symptoms and affected systems.

Table Legend

- Table 1. Occurrences and frequencies of common symptoms in filtered tweets

Part I: Text Preprocessing and Rule-based Filtering

Data collection 
471, 553, 966 Tweets from 27 months

Symptom lexicon


System	Symptom	Descriptions
Systemic	Fever	fever
Respiratory	Cough	cough
Digestive	Vomiting	vomit
Integumentary	Alopecia	hair loss

10 systems
257 symptoms

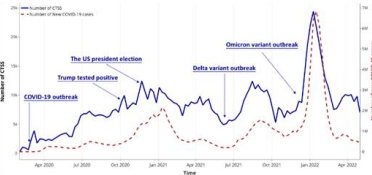
Step 1: Initially filtering of COVID-19

Step 2: Text Processing and cleaning

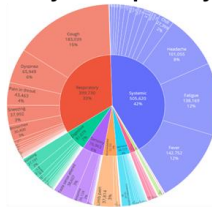
Step 3: Filtering with self-report symptoms

948, 478 Tweets with self-report symptoms 

Part II: Overall analysis of quantity and distribution



Trend in quantity of tweets and infections

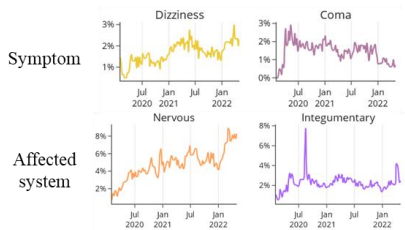


Distribution of self-reported symptoms

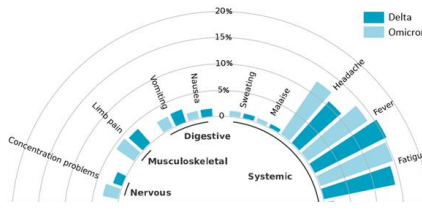
Symptom	System	DTSS Result	WHO Report
Cough	Respiratory	183039 (19.3%)	37861 (67.7%)
Fever	Systemic	142752 (15.1%)	49157 (87.9%)
Fatigue	Systemic	138169 (14.6%)	21307 (38.1%)
Headache	Systemic	101055 (10.7%)	7606 (13.6%)
Dyspnea	Respiratory	65949 (7.0%)	10402 (18.6%)
Pain in throat	Respiratory	43463 (4.6%)	7773 (13.9%)
...

Frequency of symptoms and systems

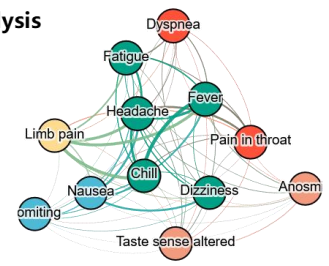
Part III: Prevalence status and comorbidity network analysis



Dynamic changes of symptom prevalence



Symptoms prevalence of different strains



The co-occurrence network analysis

Figure 1. The overall workflow

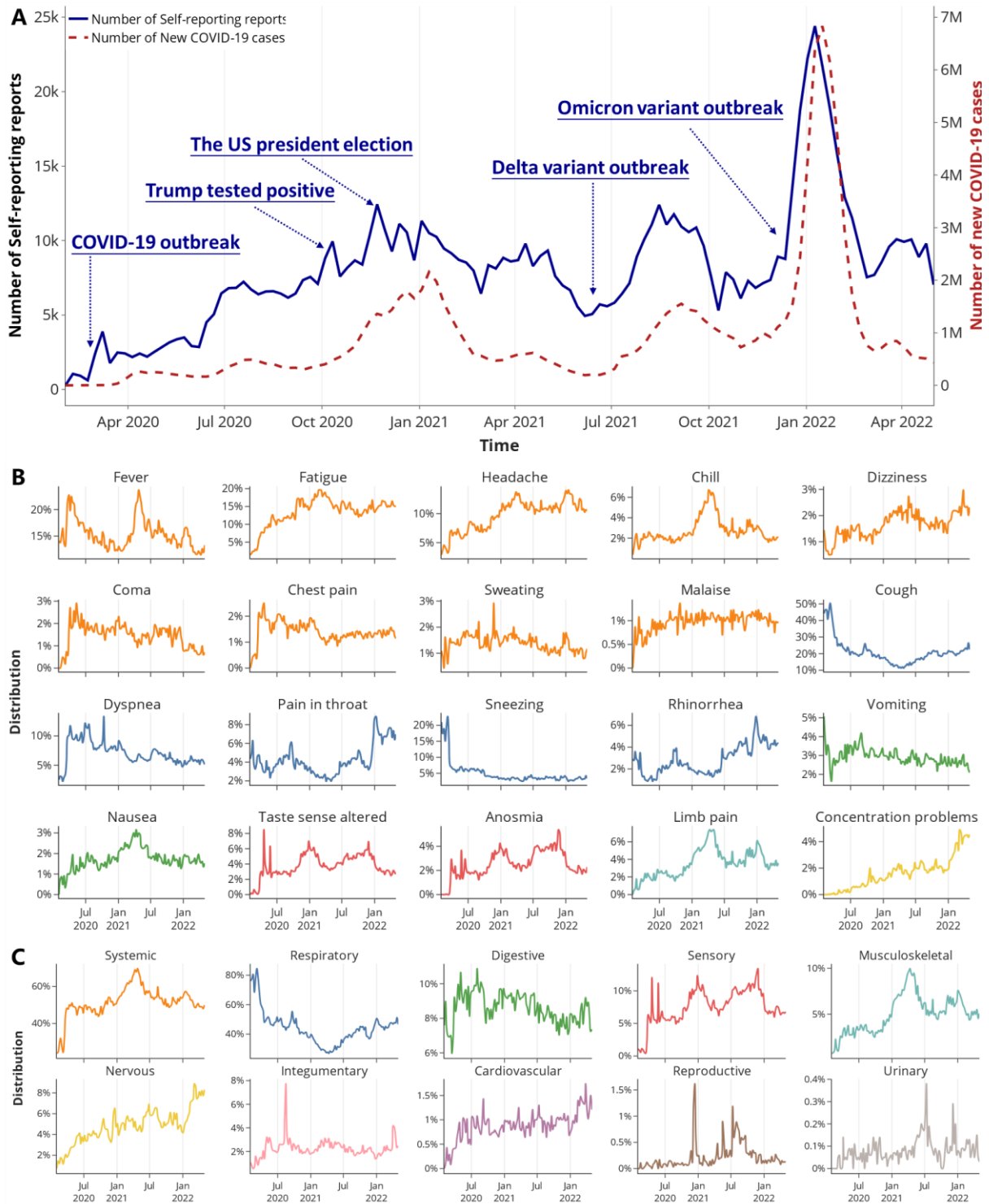


Figure 2. Weekly numbers of self-reporting tweets and weekly trends of the frequency of symptoms and affected systems (A) Weekly numbers of self-reporting COVID-19 tweets and sum of new COVID-19 cases in the US, the UK, Canada, and the Philippines. There were several waves of new cases and self-reporting tweets, including the initial outbreak in March 2020 and the

continuous rapid spread. The first peak occurred during the transition of 2020 and 2021. Weekly new cases fell back to a pre-peak level and then increased at a slow rate until the outbreak of Delta, which started a new wave of infections in middle 2021. Omicron swept across countries from December 2021, took over Delta, and gave rise to the most enormous COVID wave. During the week of January 16, 2022, weekly new cases reached the highest number of 6.83 million. The weekly self-reporting showed similar trends but with more fluctuations. Such fluctuations mainly happened with hotspot issues on social media. One example was when former US president Donald Trump tested positive for COVID during the presidential election. (B) Weekly trends of the frequency of the top 20 symptoms and (C) Weekly trends of the frequency of the affected systems. Colors of symptoms in (B) correspond to affected systems in (C).

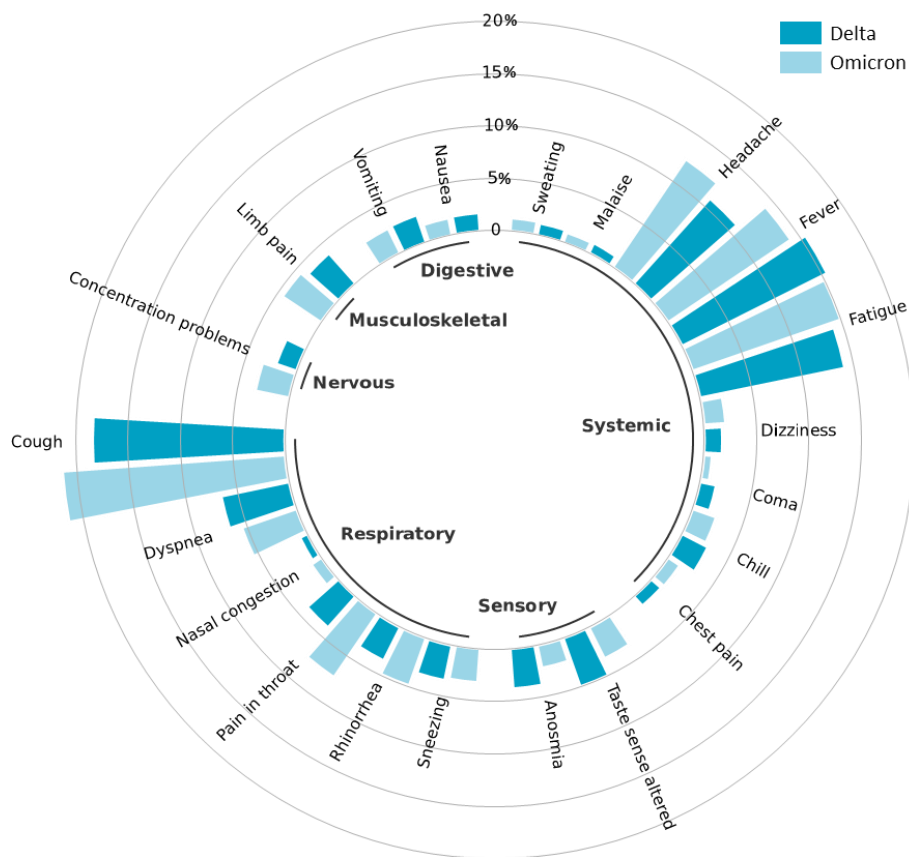


Figure 3. Distribution difference in common symptoms between Delta and Omicron

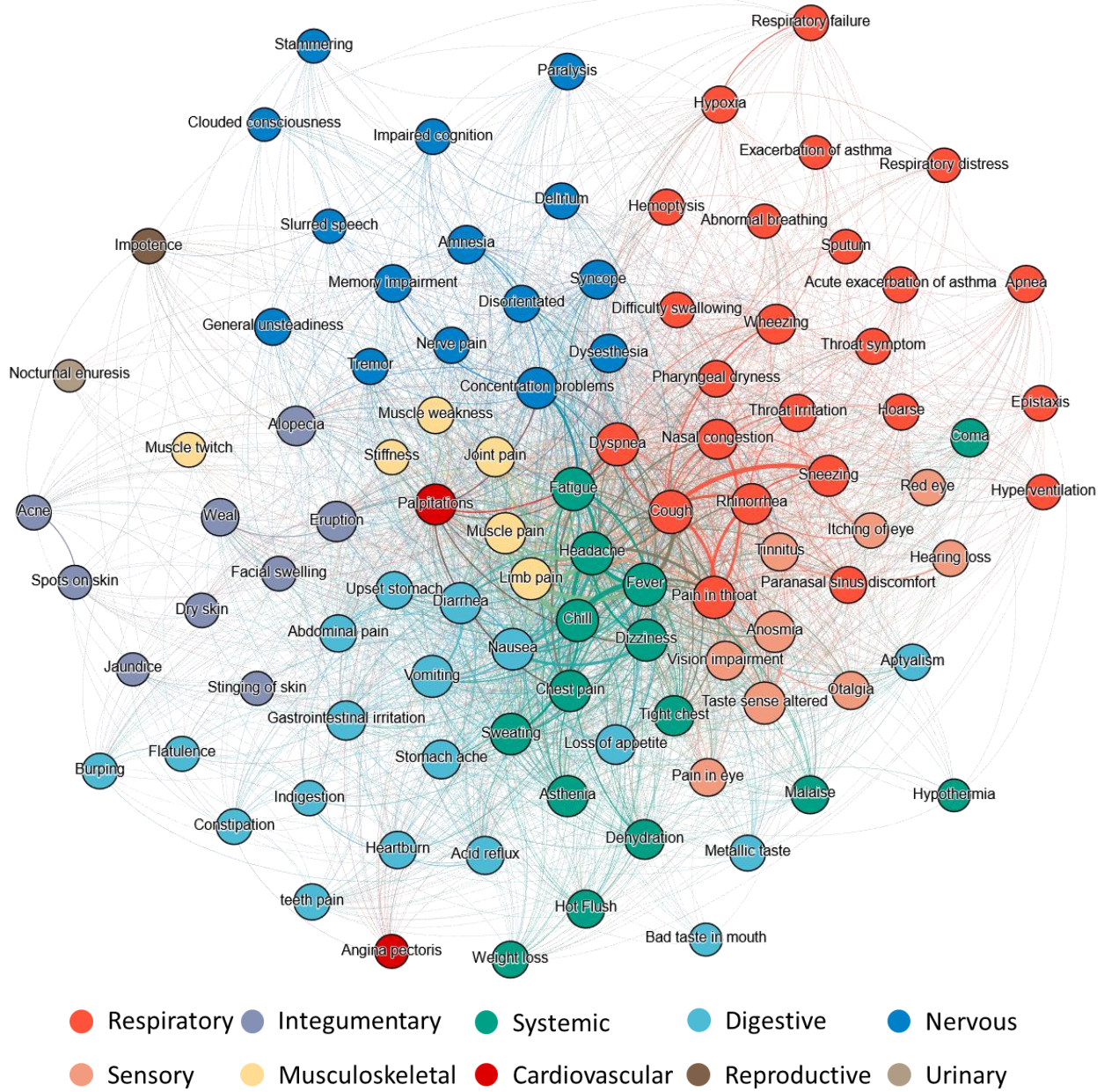


Figure 4. The co-occurrence network of different symptoms and affected systems.

Table 1. Occurrences and frequencies of common symptoms in filtered tweets

Symptoms	Body system	Self-reported (all) N=948,478	WHO [•] N=55,924	Self-reported (Delta) N=149,462	Self-reported (Omicron) N=158,994
Cough	Respiratory	183039 (19.3%)	37861 (67.7%)*	38378 (18.4%)	52325 (21.4%)
Fever	Systemic	142752 (15.1%)	49157 (87.9%)	32501 (15.5%)	34562 (14.1%)
Fatigue	Systemic	138169 (14.6%)	21307 (38.1%)	29621 (14.2%)	36704 (15.0%)
Headache	Systemic	101055 (10.7%)	7606 (13.6%)	22846 (10.9%)	30601 (12.5%)
Dyspnea	Respiratory	65949 (7.0%)	10402 (18.6%)	13841 (6.6%)	13601 (5.6%)
Pain in throat	Respiratory	43463 (4.6%)	7773 (13.9%)	8381 (4.0%)	18059 (7.4%)
Taste sense altered	Sensory	38607 (4.1%)	-	10426 (5.0%)	8188 (3.3%)
Sneezing	Respiratory	37992 (4.0%)	-	7281 (3.5%)	8024 (3.3%)
Limb pain	Musculoskeletal	37814 (4.0%)	8277 (14.8%) [▲]	8114 (3.9%)	10876 (4.4%)
Rhinorrhea	Respiratory	30400 (3.2%)	2684 (4.8%) [◆]	7570 (3.6%)	11952 (4.9%)
Chill	Systemic	27399 (2.9%)	6375 (11.4%)	5890 (2.8%)	5928 (2.4%)
Vomiting	Digestive	27191 (2.9%)	2796 (5%)*	5780 (2.8%)	6408 (2.6%)
Anosmia	Sensory	26124 (2.8%)	-	7983 (3.8%)	5525 (2.3%)
Concentration problems	Nervous	18130 (1.9%)	-	4285 (2.0%)	8104 (3.3%)
Nausea	Digestive	17238 (1.8%)	-	3675 (1.8%)	4187 (1.7%)
Dizziness	Systemic	16628 (1.8%)	-	3701 (1.8%)	5047 (2.1%)
Coma [■]	Systemic	13532 (1.4%)	-	3295 (1.6%)	2028 (0.8%)
Chest pain	Systemic	13382 (1.4%)	-	2634 (1.3%)	3312 (1.4%)
Sweating	Systemic	13255 (1.4%)	-	2511 (1.2%)	3053 (1.2%)
Malaise	Systemic	9699 (1.0%)	-	2165 (1.0%)	2573 (1.1%)
Nasal congestion [■]	Respiratory	7511 (0.8%)	-	1726 (0.8%)	2952 (1.2%)

* Specifically dry cough

[▲] Including myalgia (limb pain) and arthralgia (joint pain)

[◆] Reported as nasal congestion, including rhinorrhea and nasal congestion (count 3673, frequency 0.6%, rank 20) in self-reported symptoms.

[★] Including vomiting and nausea

[•] Reported by WHO but not the top symptoms in self-reported symptoms: hemoptysis (WHO: 503, 0.9%, ranked the 13th position, DTSS: 614, 0.1%, ranked the 75th position)

[■] For Omicron, nasal congestion reached 1.2% and replaced coma as its top 20 symptoms.