

1 **FRONT MATTER**

2
3 **Title**

4 The sexual and gender-diverse face more health challenges during COVID-19: A large-
5 scale social media analysis with natural language processing.

6
7 **Authors**

8 Zhiyun Zhang¹, Yining Hua^{2,3,4}, Peilin Zhou⁵, Shixu Lin¹, Minghui Li¹, Yujie Zhang¹, Li
9 Zhou⁴, Yanhui Liao⁶, Jie Yang^{1*}

10
11 **Affiliations**

12 ¹ Department of Big Data in Health Science School of Public Health, Zhejiang University
13 School of Medicine, Hangzhou, China.

14 ² Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston,
15 USA.

16 ³ Department of Biomedical Informatics, Harvard Medical School, Boston, USA.

17 ⁴ Division of General Internal Medicine and Primary Care, Department of Medicine,
18 Brigham and Women's Hospital, Boston, USA.

19 ⁵ Thrust of Data Science and Analytics, The Hong Kong University of Science and
20 Technology (Guangzhou), Guangzhou, China.

21 ⁶ Department of Psychiatry, Sir Run Run Shaw Hospital, Zhejiang University School of
22 Medicine, Hangzhou, China.

23
24
25 **Corresponding author**

26 Jie Yang, Ph.D.

27 School of Public Health, Zhejiang University School of Medicine, Hangzhou, China

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

29 Email: jielynlp@gmail.com

1 Abstract

2 **Background:** The COVID-19 pandemic has caused a disproportionate impact on the sex
3 and gender diversity (SGD) community. Compared with non-SGD populations, their social
4 relations and health status are more vulnerable, whereas public health data regarding SGD
5 is scarce. **Methods:** To analyze the concerns and health status of SGD individuals, this
6 cohort study leveraged 471,371,477 tweets from 251,455 SGD and 22,644,411 non-SGD
7 users, spanning from February 1, 2020, to April 30, 2022. The outcome measures comprised
8 the distribution and dynamics of COVID-related topics, attitudes towards vaccines and the
9 prevalence of symptoms. **Results:** Topic analysis revealed that SGD users engaged more
10 frequently in discussions related to “friends and family” (20.5% vs 13.1%, $P<0.001$) and
11 “wear masks” (10.1% vs 8.3%, $P<0.001$) compared to non-SGD users. Additionally, SGD
12 users exhibited a significantly higher proportion of positive sentiment in tweets about
13 vaccines, including Moderna, Pfizer, AstraZeneca, and Johnson & Johnson. Among
14 102,464 users who self-reported COVID-19 diagnoses, SGD users disclosed significantly
15 higher frequencies of mentioning 61 out of 69 COVID-related symptoms to non-SGD users,
16 encompassing both physical and mental health challenges. **Conclusion:** The results provide
17 insights to an understanding of the unique needs and experiences of the SGD community
18 during the pandemic, emphasizing the value of social media data in epidemiological and
19 public health research.

20 21 MAIN TEXT

22 1. Introduction

23 The COVID-19 pandemic has posed tremendous pressure on global health systems, leading
24 to issues such as resource constraints and overcrowding in medical facilities [1]. These
25 challenges are particularly acute for vulnerable communities, including sexual and gender
26 diverse (SGD) individuals, who frequently face systemic inequality [2]. Prior studies have
27 reported that individuals from these communities encounter increased barriers to healthcare
28 access due to both overt and systemic discrimination, as well as inadequate health insurance
29 coverage [1, 3]. They often have a higher prevalence of chronic conditions that are
30 associated with severe COVID-19 outcomes, such as diabetes [4], cardiovascular diseases,
31 and respiratory conditions like asthma [5, 6]. In addition, systemic factors also placed these
32 individuals in socially or emotionally challenging environments, heightening their risk for
33 severe mental health issues [7].

34 This backdrop of heightened vulnerability underscores the crucial role that vaccine
35 availability and acceptance play in curbing the spread of COVID-19 among SGD
36 populations [8, 9]. Existing research on vaccine hesitancy often overlooks or misrepresents
37 these individuals [10], despite the fact that clinical and social factors contribute to their
38 vaccine acceptance [11]. For instance, concerns over underlying health conditions [10], and
39 considerations of vaccine efficacy and safety [12] all shape their attitudes towards
40 vaccination. It is crucial, therefore, to target SGD individuals specifically, aiming to
41 enhance their vaccine acceptance by deeply understanding their stance.

42 However, research on the health of SGD individuals during the COVID-19 pandemic faces
43 a notable gap in both depth and breadth. Most studies rely on online surveys and
44 questionnaires [13, 14], constrained by the inherent biases of questionnaire design and the
45 prolonged intervals of data collection. Moreover, research utilizing electronic health records
46 typically focuses on specific symptoms of severe COVID-19, influenced by hospital
47 admission rates and delays in gathering SGD data [15, 16], thus lacking in generalizability.
48 Additionally, the access to electronic health records is highly restricted, limiting the
49 coverage of patients [17]. A major challenge in conducting comprehensive health status

1 analysis for SGD populations lies in selecting large-scale, representative cohorts, which
2 underscores the potential of alternative data sources, such as social media, which has been
3 increasingly validated as a valuable tool in public health research [18-20]. Social media-
4 related studies cover topics ranging from mental healthcare [21-23] to disease symptoms
5 [24] and public acceptance of treatments [25-27], both in the context of COVID-19 and
6 previous health crises such as H1N1 [28] and Zika [29]. Meanwhile, new deep learning-
7 based language models, pipelines, and datasets [30-33] offer opportunities to analyze the
8 massive textual information from social media platforms. This synergy between natural
9 language processing (NLP) and social media analytics opens up novel avenues for research
10 that span both data collection and analytical interpretation [18].

11 Within this context, our study leverages large-scale Twitter data and NLP methodologies to
12 scrutinize the health and well-being of SGD individuals during the COVID-19 pandemic.
13 We address three principal research questions: 1) What are the predominant topics discussed
14 by SGD Twitter users during the pandemic? 2) How concerned are SGD individuals about
15 pandemic precautions, such as mask-wearing and vaccination? 3) Do SGD individuals face
16 more acute symptom risks and mental health challenges compared to non-SGD individuals
17 during the pandemic?

18 To address the questions above, we employ Latent Dirichlet Allocation (LDA) models to
19 discern public discussion themes and track their temporal evolution. Named Entity
20 Recognition (NER) and Targeted Sentiment Analysis (TSA) models—both grounded in
21 advanced NLP techniques and trained on Twitter-specific datasets—are used to compare
22 vaccine perceptions between SGD and non-SGD individuals. We also identify and analyze
23 Twitter users who have self-reported a COVID-19 diagnosis to compare health outcomes
24 across SGD and non-SGD groups. Our preliminary results underline that SGD individuals
25 manifest significantly elevated symptomatology and mental health challenges, emphasizing
26 an imperative for specialized interventions.

27 **2. Methods**

28 **Experimental Design**

29 This cohort study collected a comprehensive dataset from February 2020 to April 2022,
30 with the data collection process adhering to Twitter's Terms of Service. Ethical approval
31 was secured from the Institutional Review Board of Zhejiang University and Mass General
32 Brigham. An overview of data distribution and study design is provided in **Figure S1**. SGD
33 users were identified through user profiles, and topic modeling techniques were employed
34 to analyze the content. Further statistical analyses were performed to understand their
35 sentiments regarding COVID-19 vaccines, compare self-reported symptoms between SGD
36 and non-SGD users, and investigate their mental health status.

37 **Data collection and selection**

38 This study collected tweets through leveraging tweet IDs from a public coronavirus Twitter
39 dataset [34], which follows specified accounts and collects real-time tweets mentioning
40 specific keywords. We instituted a filtering process where tweets containing URLs were
41 excluded to attenuate the impact of news and automated bot activities. Subsequently, we
42 focused on identifying tweets from SGD users, including lesbian, gay, bisexual, transgender,
43 queer, intersex, and asexual individuals [35]. SGD users were filtered through user profiles
44 using keyword filtering and regular expression matching (**Table S1**): 1) User profiles must
45 contain SGD-related keywords. 2) There should be no negation words before or after the
46 keywords. 3) The keywords should not be preceded or followed by terms such as
47 "advocator" and "supporter" as some users may advocate for SGD rights without necessarily

1 being SGD themselves. A manual validation process was conducted on a subset of 500
2 selected SGD users, achieving a classification accuracy of 93.8%. We also evaluated the
3 baseline characteristics of geographic information on the validation subset (**Table S2**).

4 **Statistical Analysis**

5 All statistical tests were conducted using Python 3.8, and were two-tailed tests, with
6 significance levels adjusted using Bonferroni corrections for multiple comparisons.

7 **Topic modeling:** Given the imbalanced dataset with a disproportionate number of tweets
8 from non-SGD users, we performed a random under-sampling to achieve parity in tweet
9 numbers (n=2,296,289 for both groups) and sensitivity analysis was applied to verify the
10 stability of the under-sampling (Supplementary methods and **Table S4**). The random
11 sampled tweets were then preprocessed through: 1) removing the mention symbol "@" and
12 the quoted usernames, 2) removing stop and short words with less than two letters, 3)
13 applying word lemmatization, 4) adding bigrams and trigrams that co-occur more than five
14 times, and 5) removing short tweets containing less than five tokens.

15 After preprocessing, 3,498,468 tweets were subjected to Latent Dirichlet Allocation (LDA)
16 [36] using the Gensim package [37]. Model selection criteria included both topic coherence
17 and model perplexity, tested over a range of 10 to 50 topics. The topic number was set to
18 12 in our case according to experiments on balancing coherence and perplexity scores
19 (**Figure S2**). Topic validity was further confirmed through visualization using pyLDAvis
20 [38] and manual inspection of the top 20 keywords per topic (**Table S3**). To compare the
21 discussion differences between SGD users and non-SGD users on a specific topic, we
22 applied *Scattertext* [39] to visualize the word frequency.

23 **Sentiment Analysis of vaccines:** We used a pre-trained language model, COVID-Twitter-
24 BERT [30], which was a BERT-LARGE structure pre-trained on 160 million COVID-19-
25 related tweets, as the backbone for our named entity analysis (NER) and target sentiment
26 analysis (TSA) models. A linear layer and a SoftMax function were added to the end of CT-
27 BERT [40] to predict the span of each vaccine entity for the NER model. The encoder of
28 BERT-SPC [41] was replaced with CT-BERT for the TSA model. The models were fine-
29 tuned on the training set of the Medical Entities and Targeted Sentiments on COVID-19-
30 related tweets (METS-CoV) dataset [32] using NCRF++ [42], both are part of our prior
31 work. This dataset included annotations for vaccine entities and their corresponding
32 sentiment labels in tweets. The performance of the NER and TSA models was tested on the
33 vaccine entity from the METS-CoV test set and resulted in an F1 score of 90.44% and an
34 accuracy of 79.15%, respectively. As most of the recognized vaccine entities were informal
35 expressions or misspelled, we manually incorporated the expressions of vaccine entities
36 (details provided in **Table S5**) and selected the four most frequently mentioned COVID-19
37 vaccines (Moderna, Pfizer, AstraZeneca, and Johnson & Johnson) for in-depth analysis.

38 **Symptom extraction and identification:** Tweets that self-reported COVID-19 diagnoses
39 were identified using lexicon filtering (details provided in **Table S6**). Tweets not written in
40 the first person or contained negative or uncertain expressions (e.g., wonder, thought, might,
41 etc.) before and after the keywords were filtered out through regular matching. For each
42 selected user, we determined the date of diagnosis based on the content of the first self-
43 report tweet. If the date of diagnosis was not specified in that tweet, we assumed that the
44 time of tweeting was the time of diagnosis.

45 Tweets posted before and after 30 days of their self-report date were collected for users who
46 self-reported COVID-19 diagnoses. We then screened these tweets using a COVID-19

1 symptom lexicon developed by Wu et al. [43], which contains commonly used synonyms
2 and colloquial variants on social media that pertain to symptoms and their associated
3 affected organs or systems. For the identification of mental health-related tweets, we utilized
4 an exhaustive mental health lexicon [44] which has been rigorously validated by
5 professionals in the fields of psychiatry and psychology. This lexicon encompasses 231
6 keywords distributed across four major mental health conditions: anxiety, depression,
7 insomnia, and substance use disorders.

8 9 **3. Results**

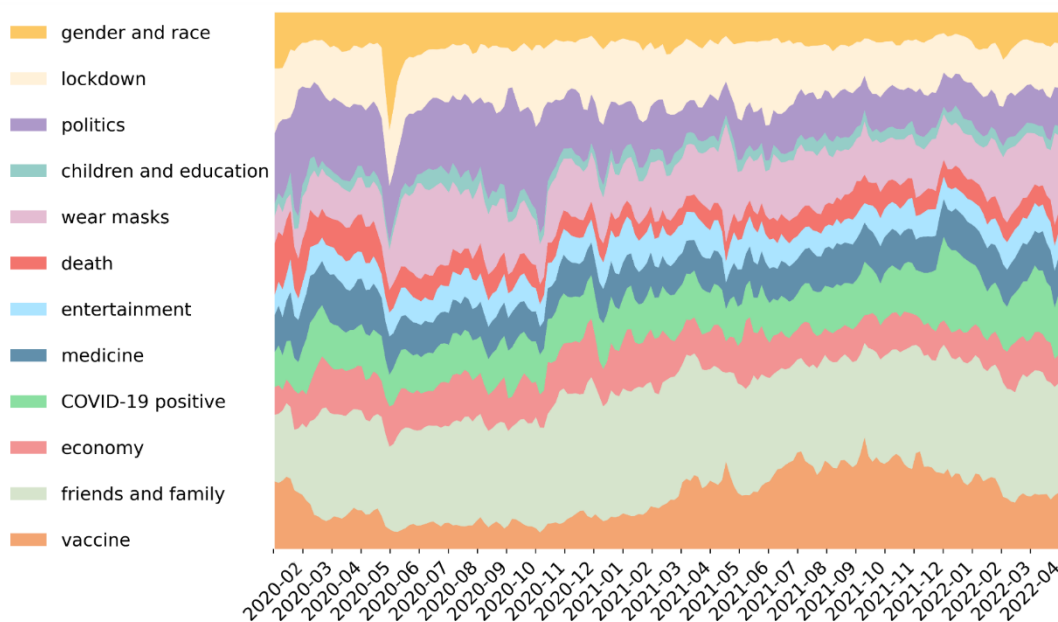
10 We downloaded a total of 471,371,477 tweets from the public COVID-19 Twitter dataset .
11 After excluding tweets with URLs, the dataset was narrowed down to 169,669,346 tweets.
12 Within this set, 2,296,289 tweets originated from 251,455 SGD users and 167,373,057
13 tweets originated from 22,644,411 non-SGD users (details provided in **Table S1**).

14 **Topic distribution and discrepancies**

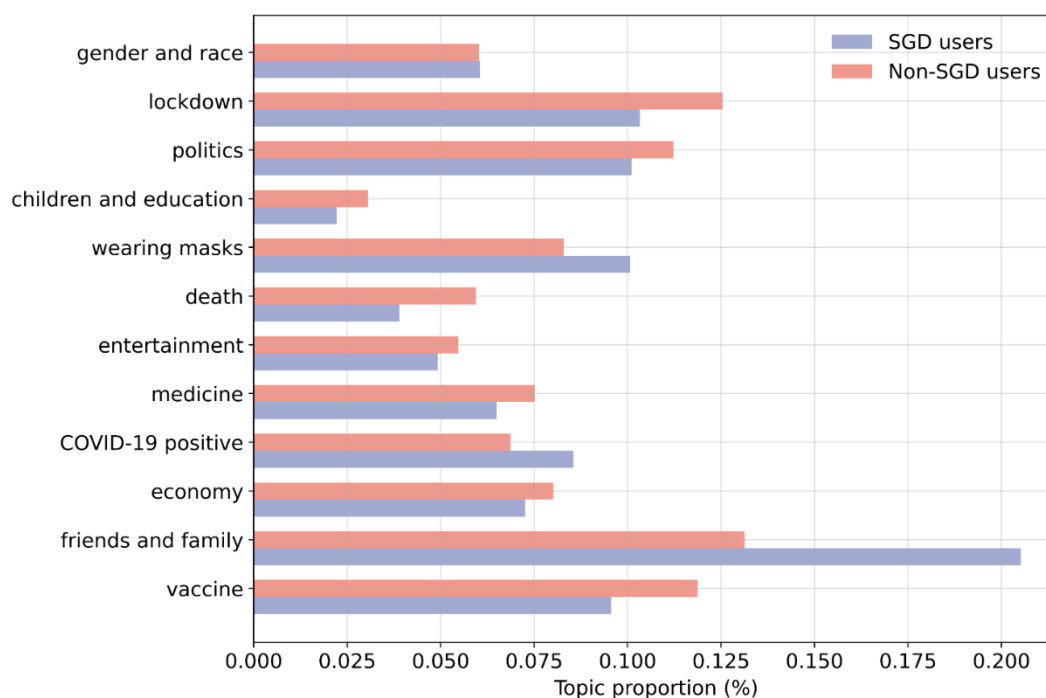
15 After preprocessing, 3,498,468 tweets were subjected to the topic model, and the top 5 most
16 discussed topics were "friends and family" (16.8%, 95% CI - 16.7% to 16.9%), "lockdown"
17 (11.4%, 95% CI - 11.3% to 11.5%), "vaccine" (10.7%, 95% CI - 10.6% to 10.8%),
18 "politics" (10.7%, 95% CI - 10.6% to 10.8%), and "wearing masks" (9.2%, 95% CI - 9.1%
19 to 9.3%). Notably, these trends varied over time and exhibited specific peaks, as smoothed
20 by a 7-day moving average (**Figure 1A**). Besides, topic fluctuations reflect significant
21 events during the pandemic. For instance, the sharp increase in discussions on "gender and
22 race" in May 2020 corresponds to heightened concerns about racial equity [45] following
23 the police killing of George Floyd, an unarmed Black civilian [46]. Similarly, the U.S.
24 presidential election and subsequent debates over COVID-19 policies [47] caused a spike
25 in the topic "politics" in November 2020.

26 When comparing topic frequencies between SGD and non-SGD users using χ^2 test with
27 Bonferroni adjusted significance level $P < 0.004$, the data indicate that discussions about
28 "friends and family" were significantly more prevalent among the former (20.5% vs 13.1%,
29 $P < 0.001$) (**Figure 1B**). For a deeper insight into the discussion variations, we used
30 *Scattertext* [39] to visualize the word frequencies (**Figure S3**). The result highlights that
31 terms regarding family members occur frequently, with SGD users often mentioning
32 "partner" while non-SGD users more frequently use terms like "daughter" and "baby".
33 Furthermore, SGD users express a range of emotions more frequently, especially negative
34 ones like "anxiety", "upset", "angry", "depression". For other topics, SGD users were more
35 likely to talk about "wear masks" (10.1% vs 8.3%, $P < 0.001$) and "COVID-19 positive"
36 (8.6% vs 6.9%, $P < 0.001$), while non-SGD users discussed other topics like "vaccine" (9.6%
37 vs 11.9%, $P < 0.001$) and "lockdown" (10.3% vs 12.5%, $P < 0.001$) more often.

38 (A)



1 (B)



2 **Fig. 1.** Topic distributions in COVID-19-related tweets. (A) The topic proportion
3 distribution of SGD user-posted tweets over time. (B) Topic distributions of tweets posted
4 by SGD users and non-SGD users.

5 Attitude Towards COVID-19 Vaccines

6 **Table 1** shows the distribution of three-category sentiments (positive, neutral, and negative),
7 wherein the majority were characterized as neutral. Notably, sentiments toward the Pfizer
8 vaccine exhibited the highest frequency of positive evaluations. Contrarily, attitudes toward
9 the AstraZeneca vaccine appeared the most polarized among SGD users when contrasted
10 with non-SGD users. Utilizing an Independent Samples t-test for statistical analysis, we
11 found that SGD users displayed significantly higher proportions of positive sentiments for
12 all four vaccine types including Moderna (tweets no. [%]. 610 [12.6] vs 28828 [8.6],

P<0.01), Pfizer (tweets no. [%]. 984 [12.9] vs 58109 [8.8], P<0.01), AstraZeneca (tweets no. [%]. 158 [11.0] vs 16227 [3.8], P<0.01), and Johnson & Johnson (tweets no. [%]. 208 [9.3] vs 11482 [6.5], P<0.01). Furthermore, the proportions of negative sentiments for both the Moderna (tweets no. [%]. 182 [3.7] vs 21271 [6.4], P<0.01) and Pfizer (tweets no. [%]. 327 [4.3] vs 51669 [7.8], P<0.01) vaccines were significantly lower among SGD users compared to the non-SGD group.

Physical and mental health status

We identified 2,098 SGD and 100,366 non-SGD users who self-reported COVID-19 diagnoses (the overview of users' filtering process is provided in **Figure S4**). Analysis of tweets within a 30-day window surrounding the self-reported date yielded mentions of 69 unique symptoms, implicating 15 distinct organ systems or physiological functions (**Supplementary File 1**). An independent samples t-test showed that the frequency of mentions for 61 of these 69 symptoms was significantly higher ($P < 7.25 \times 10^{-4}$, Bonferroni adjusted) among the SGD cohort compared to the non-SGD group.

Table 1. Percentage of positive and negative sentiments in vaccine-related tweets.

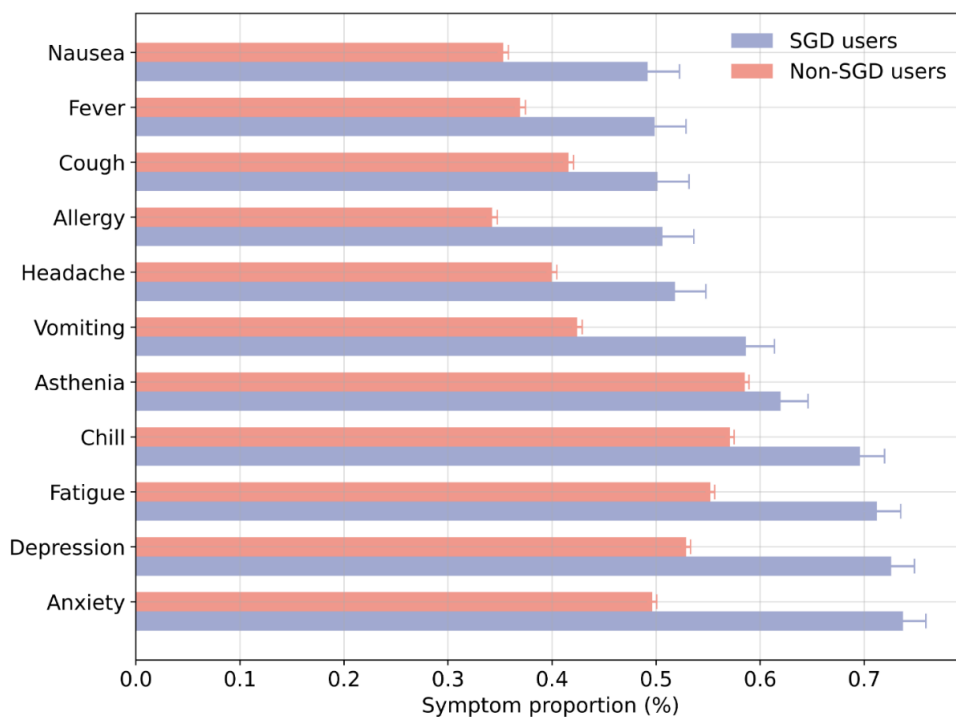
Vaccine	Positive			Negative		
	Tweets, No. (%) [95% CI ^a]		P value ^b (SGD greater)	Tweets, No. (%) [95% CI]		P value (SGD smaller)
	SGD	Non-SGD		SGD	Non-SGD	
Moderna	610 (12.6) [11.6-13.5]	28828 (8.6) [8.5-8.7]	<0.001	182 (3.7) [3.2-4.3]	21271 (6.4) [6.3-6.5]	<0.001
Pfizer	984 (12.9) [12.2-13.7]	58109 (8.8) [8.7-8.8]	<0.001	327 (4.3) [3.8-4.8]	51669 (7.8) [7.7-7.9]	<0.001
AstraZeneca	158 (11.0) [9.4-12.7]	16227 (3.8) [3.8-3.9]	<0.001	68 (4.8) [3.7-5.9]	11071 (2.6) [2.6-2.7]	0.999
Johnson & Johnson	208 (9.3) [8.1-10.5]	11482 (6.5) [6.4-6.6]	<0.001	127 (5.6) [4.7-6.6]	11191 (6.3) [6.2-6.4]	0.101

^a We used the ratio t-test to calculate 95% confidence intervals and the independent samples t-test to calculate significance.

^b Significance were set at $P < .006$ after Bonferroni correction.

We then calculated mention rates for each symptom in both groups. **Figure 2A** displays mention rates for symptoms cited by more than 1,000 individuals in the SGD group and 35,000 in the non-SGD group. Symptoms most frequently mentioned—such as anxiety, nausea, and allergic reactions—had higher prevalence among SGD individuals. **Figure 2B** further shows that mention rates for symptoms related to mental and musculoskeletal health were especially elevated in the SGD individuals, followed by mental symptoms, trachea and lung, and brain.

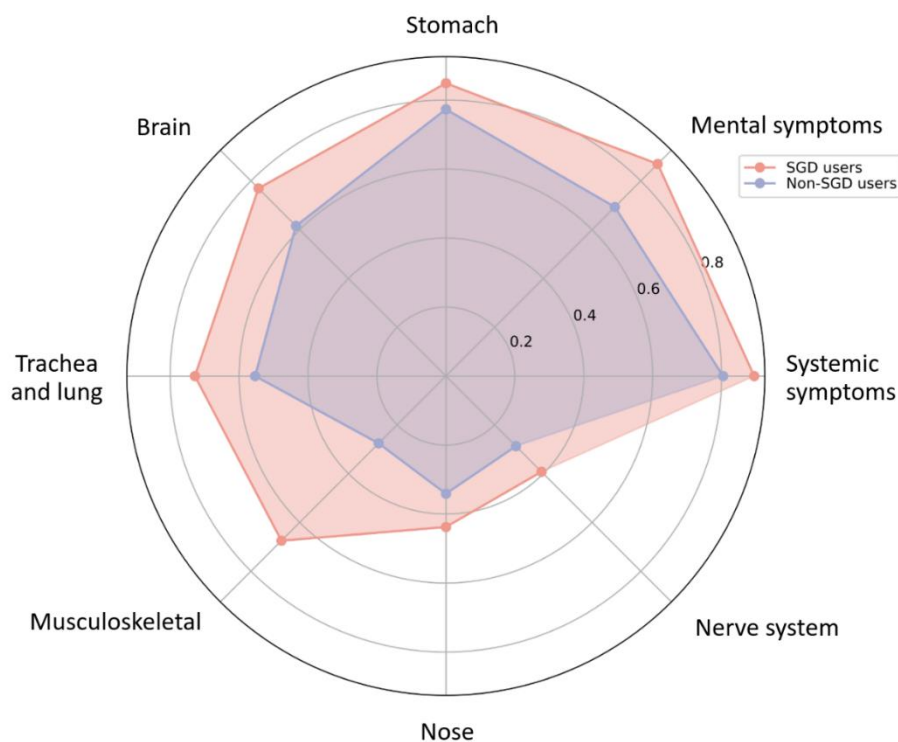
(A)



1

2

(B)



3

4

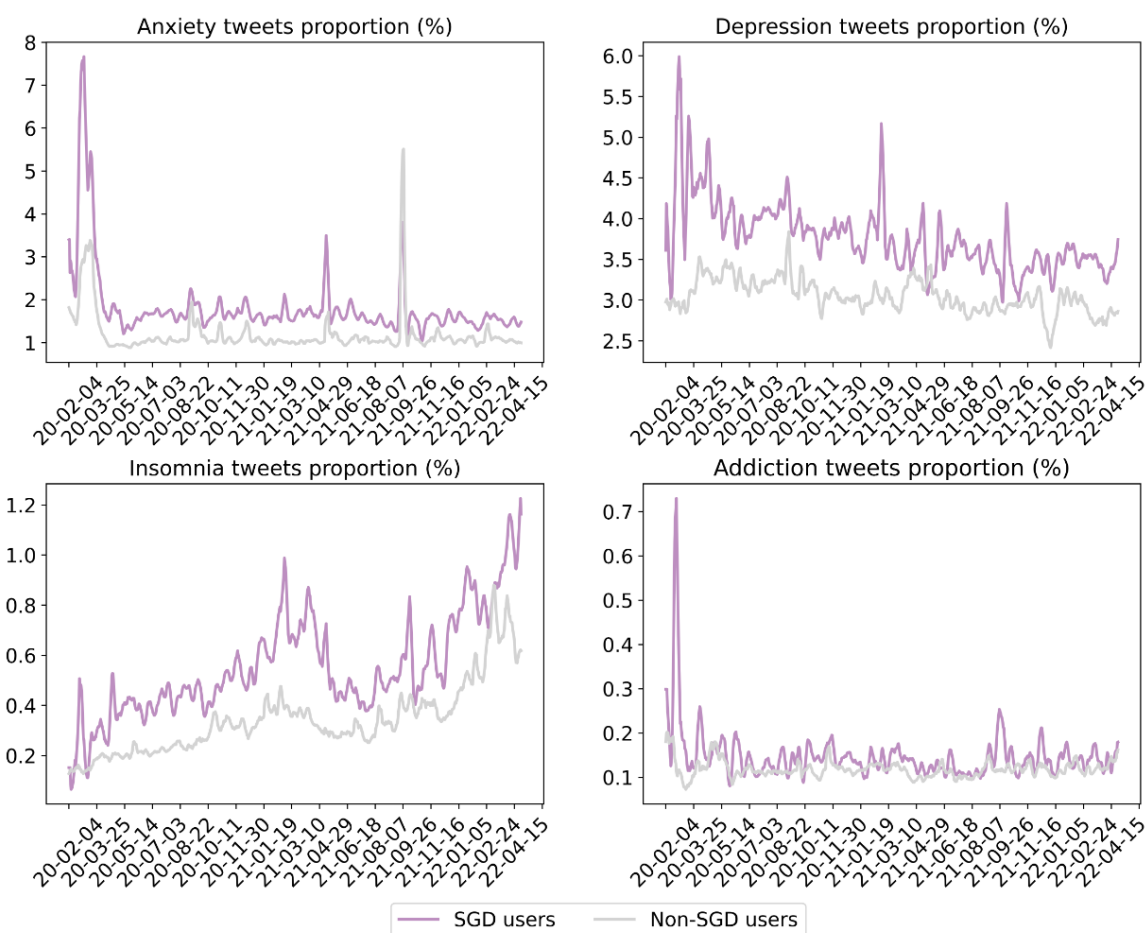
5

6

Fig. 2. Symptom distributions in users who self-report COVID-19 positive. (A) Symptom distributions of SGD and non-SGD Twitter users. (B) Distribution of the eight most affected organs or systems of SGD and non-SGD Twitter users.

1 For mental health analysis, from 137,860 mental health-related tweets contributed by SGD
 2 individuals and 7,647,024 tweets contributed by non-SGD individuals, we identified
 3 1,984,317 tweets related to anxiety, 5,258,324 related to depression, 586,648 related to
 4 insomnia, and 200,905 related to substance use disorders.

5 **Figure 3** shows the temporal distribution of these mental health-related tweets. We
 6 observed an initial surge in tweets concerning anxiety, depression, and substance use at the
 7 onset of the pandemic, followed by a stabilization to baseline levels. Contrastingly, the
 8 proportion of tweets pertaining to insomnia demonstrated a continual increase over time.
 9 Given that the tweet distribution over time deviated from normality, we employed two-
 10 tailed Wilcoxon matched-pairs signed-ranks tests (**Table 2**), and it turned out that SGD
 11 users exhibited higher prevalence of anxiety (% tweets median [IQR]. 1.58 [0.53] vs 1.05
 12 [0.16], $P < 0.01$), depression (% tweets median [IQR]. 3.63 [0.89] vs 3.02 [0.31], $P < 0.01$),
 13 insomnia (% tweets median [IQR]. 0.52 [0.33] vs 0.32 [0.15], $P < 0.01$), and addiction (%
 14 tweets median [IQR]. 0.13 [0.10] vs 0.12 [0.02], $P < 0.01$) symptoms.



15 **Fig. 3.** The proportion of tweets concerning mental health to the total number of tweets. We
 16 used a 7-day moving average to smooth the curve.
 17

18 **Table 2.** The daily proportion of mental health-related tweets of SGD and non-SGD users.
 19 W represents the sum of the ranks of the differences above zero.

Mental symptoms	SGD (% tweets), median(IQR ^a)	Non-SGD (% tweets), median (IQR)	W	P-value ^b
anxiety	1.585 (0.530)	1.047 (0.158)	317574	<0.001
depression	3.634 (0.885)	3.024 (0.315)	302044	<0.001

insomnia	0.523 (0.325)	0.320 (0.154)	312525	<0.001
addiction	0.127 (0.103)	0.115 (0.024)	203679	<0.001

^a We used the IQR and Wilcoxon matched-pairs signed-ranks test in the statistical analysis.

^b Significance were set at $P < .016$ after Bonferroni correction.

4. Discussion

In this pioneering social media-based retrospective cohort study, we examined the differential impact of the COVID-19 pandemic on SGD individuals by analyzing a large corpus of pandemic-related tweets over two and a half years. Our methodological approach encompasses: 1) the use of topic modeling to delineate concerns unique to SGD individuals; 2) the application of advanced deep learning-based NLP algorithms for sentiment analysis towards vaccines; and 3) a comparison of self-reported COVID-19 symptoms between SGD and non-SGD individuals. Through these avenues, our research aims to elucidate the unique challenges confronting SGD individuals during the pandemic and to inform targeted interventions designed to alleviate their physical and psychological burdens, and therefore enhance their well-being.

Our topic modelling results divulge a heightened focus among SGD individuals on themes like "friends and family", where they tend to express negative emotions more frequently. Research indicates that pandemic policies, such as school closures and lockdowns, have severed some social connections, leaving SGD individuals more dependent on family members [3]. However, older individuals are often isolated as they are four times less likely to have children and SGD youth are forced to be at home with unsupportive parents [7]. Emotional and mental health harms may arise from the lack of supportive surroundings. We also observed that SGD individuals are more likely to discuss topics related to preventative health measures such as "mask-wearing" and "COVID-19 testing." These observations align with previous research by Sears Brad et al. conducted in the US [48]. They found that SGD individuals prefer to wear masks (94.0% vs. 89.9%) and take COVID-19 tests (38.3% vs. 29.0%).

Our sentiment analysis further shows more positive attitudes towards vaccines within SGD individuals compared to their non-SGD counterparts, which is positively correlated with stronger vaccination stance [49, 50]. This higher rate of vaccine adoption is corroborated by telephonic surveys and suggests greater compliance with public health directives within these communities [51, 52]. These results suggest that SGD individuals exhibit a greater awareness of the importance of precautionary measures and prefer to comply with public health orders during the pandemic [53]. The higher willingness of SGD individuals to vaccinate may be linked to their altruistic tendencies [54] and higher levels of perceived health vulnerability [55]. Nonetheless, vaccination rates for SGD populations vary widely over different regions and ethnicities [10]. More efforts are needed to assess vaccination rates in these areas and improve the coverage for SGD individuals without insurance or documents.

In examining self-reported symptoms among COVID-19-positive individuals, our data reveals concordance with clinical studies [56-58] regarding the most frequently mentioned symptoms. However, we found a higher frequency of certain mild symptoms such as musculoskeletal and mental health issues within the SGD population, as compared to electronic health record-based studies [59]. This underscores the utility of social media as a complementary data source for capturing a broader spectrum of patient experiences that might not be adequately recorded in clinical settings. In addition, we noted that SGD

1 individuals are more likely to experience more severe symptoms after COVID-19 infection.
2 This may be due to inadequate health insurance coverage and higher-than-average rates of
3 underlying diseases such as diabetes, and asthma, which can increase the risk of severe
4 symptoms [5, 6]. Besides, SGD users mentioned musculoskeletal symptoms (body pain,
5 myalgia pain, arthralgia pain, etc.) at a particularly higher rate compared to non-SGD users.
6 These symptoms are often associated with severe disease as they can be triggered by
7 increased inflammatory factors (e.g., interleukin-6) during infection [60, 61].

8 The psychological ramifications within SGD communities warrant nuanced attention, as our
9 study indicates elevated rates of mental health symptoms than non-SGD groups during the
10 pandemic, which is consistent with pre-existing literature employing the PHQ4 scale and
11 online surveys [13, 14, 62]. In terms of temporal variations, the frequency of insomnia-
12 related tweets exhibited a correlation with diagnosed COVID-19 cases, peaking in January
13 2021 and rising steadily from January through April 2022. These trends are congruent with
14 clinical literature suggesting a high correlation between insomnia and COVID-19 infection
15 [63-65]. Conversely, fluctuations in tweets related to other mental health conditions—
16 namely depression, anxiety, and addiction—appeared to be more significantly influenced
17 by social determinants. For instance, a sudden spike in anxiety-related tweets occurred in
18 September 2021, and the majority of the discussion was focused on the increase in fuel
19 prices. This phenomenon has also been observed by previous research [44]. During the
20 pandemic's initial outbreak in February 2020, SGD communities experienced more
21 pronounced spikes in symptoms of depression, anxiety, and addiction compared to their
22 non-SGD counterparts. These exacerbated symptoms may be attributed to distinct and more
23 severe social challenges confronting SGD individuals, such as limited access to supplies
24 and healthcare [66]. Moreover, it is crucial to highlight the constrained social support
25 networks often associated with SGD communities, which include family, partners, and peers.
26 Such networks frequently lack the resilience and social capital to act as effective buffers
27 against the immediate repercussions of both social changes and health crises [67].

28 We acknowledge several limitations in our study. First, the age and geographical
29 distributions of Twitter users are skewed, introducing potential selection bias that may limit
30 the external validity of our findings. For instance, individuals with lower socio-economic
31 status or those of advanced age may be underrepresented on Twitter, thereby introducing a
32 bias towards certain demographic groups [68]. Secondly, despite the application of
33 advanced natural language processing models employing deep learning, our pipeline is
34 susceptible to misclassification bias due to lexical ambiguity [69]. To assess this issue, we
35 conducted a random selection of 500 tweets identified as originating from SGD individuals.
36 Manual validation of these tweets suggests that 31 (6.2%) tweets may have been
37 inaccurately categorized. Moreover, the composition of our non-SGD control group is
38 subject to information bias; SGD users who have not publicly disclosed their identities on
39 Twitter might be included, which could attenuate the observed effect sizes and affect the
40 internal validity of our findings. Furthermore, our study is confined by the absence of pre-
41 pandemic baseline data, largely due to Twitter's data-sharing constraints. This results in a
42 lack of temporal control, making it challenging to differentiate the health disparities
43 between SGD and non-SGD groups directly attributable to the pandemic. Strict filtering
44 criteria for users that self-reported COVID-19 positive may lead to a lower recall rate,
45 resulting in selection bias among the remaining samples and an inability to represent the
46 entire population. Moreover, the collected symptom descriptions may be subjective to the
47 user and lack the strictness of evidence-based medicine, but they can serve as an auxiliary
48 tool for public health analysis. These limitations should underscore the need for cautious
49 interpretation.

5. Conclusion

In summary, this pioneering study employs various NLP techniques like NER, TSA and LDA models, to provide an in-depth understanding of the experiences and health outcomes of SGD individuals during the COVID-19 pandemic. Our findings emphasize the importance of enhancing social and legal support for SGD individuals and informing public health interventions to address disparities during challenging times. The methodology and pipeline developed in this study can be applied to monitor the health of other populations, providing data-driven insights for more comprehensive public health services.

Ethical Approval: IRB approved by School of Public Health, Zhejiang University
ZGL202201-2

Data Availability: The codes used in this study can be accessed at
<https://github.com/zoay-zzy/COVID-twitter-SGD-analysis>.

Funding: None

Author contributions: Conceptualization: J.Y., Z.Z., and Y.L. Methodology: Z.Z., S.L., and Y.Z. Investigation: Z.Z., J.Y., and M.L. Writing—original draft: Z.Z. Writing—review & editing: J.Y., Y.L., P.Z., L.Z.

Z.Z. takes responsibility for the integrity of the work.

Conflict of interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments: none

Supplementary Materials

Methods

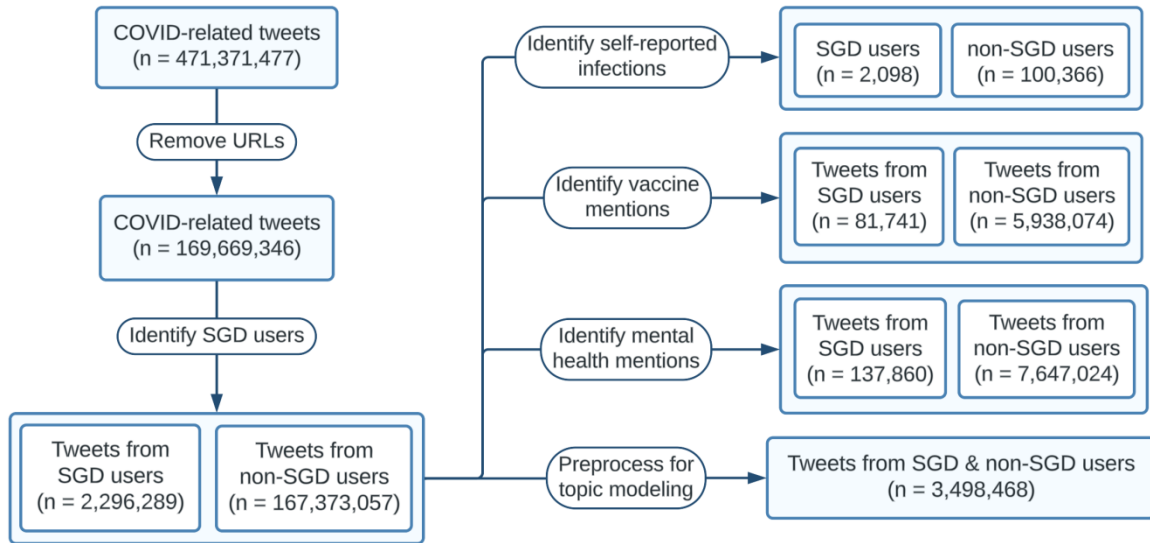
Figures S1 to S4

Tables S1 to S6

1 **Methods**

2 **Sensitivity analysis of under-sampled non-SGD tweets in the topic model.** To assess the
3 impact of different random seeds on topic distribution, we created two new samples of non-
4 SGD tweets: Sample A, with the same number as the original sample ($n = 2,296,289$), and
5 Sample B, with twice that number ($n = 4,592,578$). After preprocessing, 1,742,444 tweets
6 remained in Sample A and 3,484,880 in Sample B. Then we evaluated the distribution of
7 topics in these new corpora using the trained LDA model and compared it with the old
8 distribution using chi square test (**Table S4**).

9

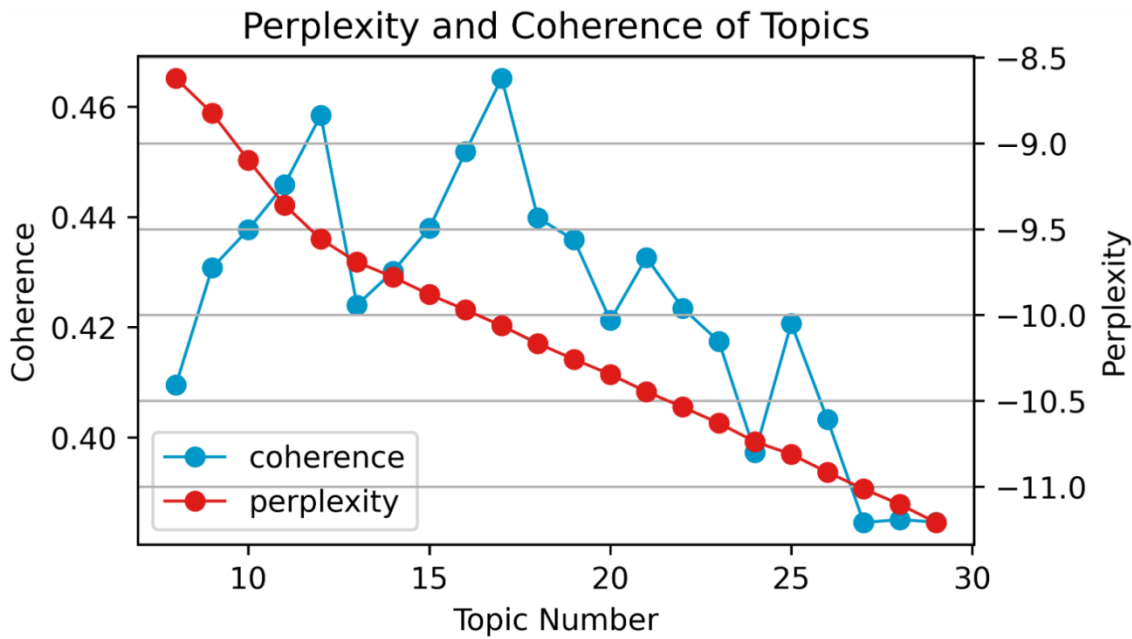


1

2

3

Figure S1. Data collection and distribution.



4

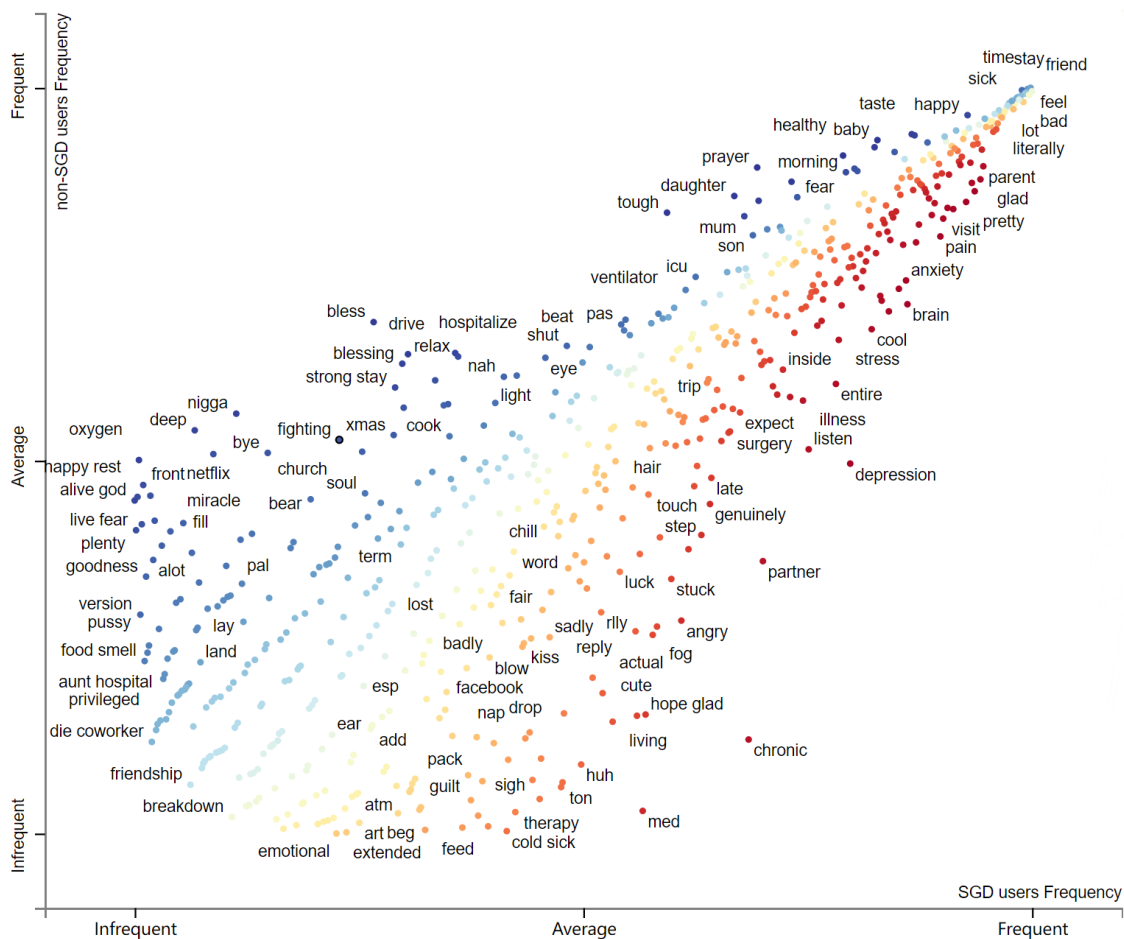
5

6

7

8

Figure S2. Perplexity and coherence score of different topics. Topic perplexity measures the generalizability of models to unseen data, and coherence measures the degree of semantic similarity between high-scoring words in topics.



1

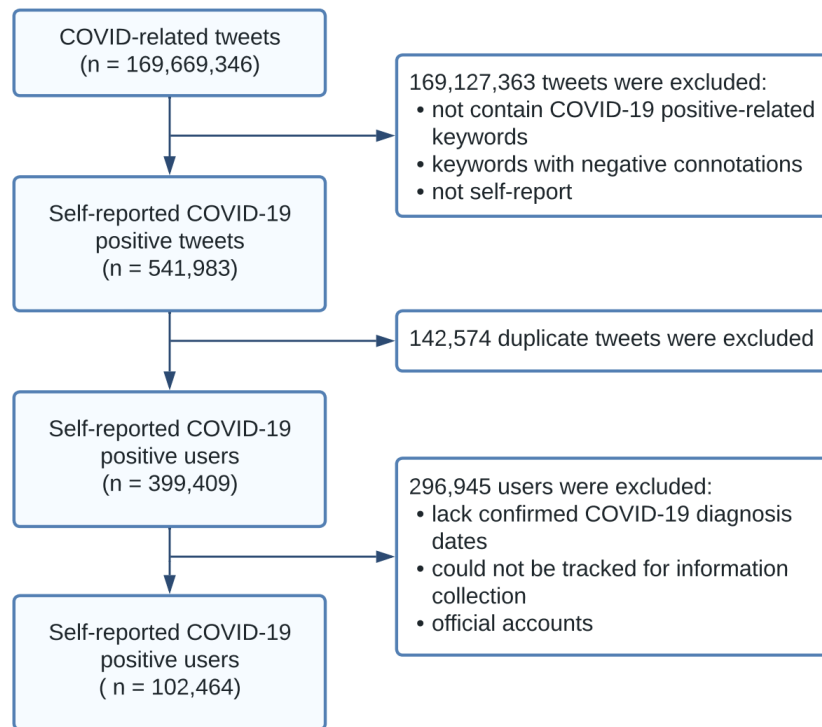
2

3

4

5

Figure S3. Visualization of word frequency in the topic “friend and family” using *Scattertext*. The x- and y- axes of terms are the dense ranks of their usage by SGD and non-SGD users respectively.



1

2

3

Figure S4. Self-reported COVID-19 positive users filtering process.

1 **Table S1.** Lexicon and rules for SGD selection. We used regular expressions to remove
 2 tweets with stop words before and after the SGD-related keywords to filter out negative
 3 expressions and SGD supporters but non-SGD individuals. "|" is used as a delimiter between
 4 lexicons.

Stop words in front	SGD lexicon	Stop words behind
not no none anti hate against pro advocate	lgbt lesbian gay bisexual transgender queer intersex asexual gender minority sex minority sex and gender minority	anti ally supporter advocator friendly

5
 6 **Table S2.** Geographic distribution of SGD and non-SGD users on the validation subset. The
 7 geographic information is extracted from the "place" field in tweet metadata but not the
 8 "location" field in "user" description, since the former is generated from GPS data.

	SGD users (%) (n=500)	non-SGD users (%) (n=500)	P value*
Country			0.015
United States	295 (59.0)	272 (54.4)	
United Kingdom	102 (20.4)	98 (19.6)	
Canada	24 (4.8)	22 (4.4)	
Australia	19 (3.8)	18 (3.6)	
India	60 (12.0)	90 (18.0)	

9 *P values were calculated using χ^2 test.

10
 11 **Table S3.** 12 topics given by the LDA model and associated top-20 keywords.

Topics	(value) Keywords	Example
vaccine	(0.081) vaccine, (0.022) vaccinate, (0.017) spread, (0.016) flu, (0.01 3) risk, (0.013) cdc, (0.012) vaccin ation, (0.012) die, (0.010) variant, (0.010) disease, (0.009) vaccinate d, (0.009) prevent, (0.009) booster, (0.008) shoot, (0.008) shot, (0.00 7) catch, (0.007) science, (0.007) i mmune, (0.006) chance, (0.006) v ax	"I see the occasional (or should that be odd?) posts by folk saying they will refuse a covid vaccine. terribly selfish. but fine as I am sure they will refuse treatment if they become ill. I do like to see darwinian natural selection in play." "It's proven that isn't true lol. you can catch covid, build antibodies, and still catch covid again because of variants. vaccines help build antibodies, thus creating stronger immunity, and when multiple people strengthen immunity through vaccination, herd immunity can happen."
friends and family	(0.026) feel, (0.022) life, (0.020) fr iend, (0.020) time, (0.018) love, (0.016) family, (0.014) bad, (0.01 4) stay, (0.014) shit, (0.012) day, (0.011) live, (0.011) die, (0.010) h ope, (0.010) lose, (0.008) mom, (0.	"Please prioritize family sponsorships as in these hard times people need their loved ones around. being alone was never easy but hadn't been so hard until these covid times."

	008) lot, (0.008) start, (0.007) hard, (0.007) god, (0.007) month	“I have no friends and havnt spoken to anyone in months coz I'd covid and I listen to your baking videos in the background to fill the void, thanks.”
economy	(0.033) job, (0.026) pay, (0.021) money, (0.015) bill, (0.014) relief, (0.010) business, (0.009) food, (0.009) lose, (0.008) cut, (0.008) company, (0.008) buy, (0.008) time, (0.007) tax, (0.007) check, (0.007) support, (0.006) cost, (0.006) struggle, (0.006) unemployment, (0.006) government, (0.005) month	“It doesn't make sense to extend covid stimulus checks to those who are working; have not lost their jobs/businesses. the checks should target unemployed people who have lost their jobs; businesses to covid. state local govts must be given money to augment their lost revenue.” “lol my landlord has been paid up to this month quit your free ride bs. my landlord has told me even if I got a job today it means nothing he is moving to evict. all because I was 25 dollars short 25 dollars short when the pandemic hit and lost my job. I got him their money later.”
covid positive	(0.073) test, (0.047) day, (0.031) week, (0.025) positive, (0.017) hour, (0.015) symptom, (0.013) time, (0.013) negative, (0.013) test_positive, (0.012) sick, (0.010) wait, (0.010) tomorrow, (0.008) appointment, (0.008) quarantine, (0.008) testing, (0.008) call, (0.007) leave, (0.007) house, (0.007) feel, (0.007) stay	“My brother got tested positive for covid almost two half weeks back, but remained asymptomatic. alhamdulillah. later my sister-in-law also tested positive with mild symptoms in quarantine at home. as a precaution, I'm going to get myself tested as well. hope it's a negative.” “After possibly being exposed to covid + me having mild symptoms, we got tested for the first time today. results expected in 24-48 hours. hopefully we're both negative”
medicine	(0.039) health, (0.027) care, (0.020) issue, (0.018) patient, (0.016) public, (0.016) hospital, (0.014) medical, (0.011) worker, (0.010) mental, (0.009) doctor, (0.009) nurse, (0.009) healthcare, (0.008) system, (0.008) staff, (0.008) community, (0.007) risk, (0.006) safety, (0.006) public_health, (0.006) service, (0.006) access	“You must work with nurses and other health care workers to make testing guidelines that ensure safety for frontline workers and patients. with regular testing now!” “The pandemic isn't over yet. all 50 states are seeing an increase in cases as we enter a potential fourth wave. hospital admissions have also climbed about 36% with deaths up 26%. vaccination efforts continue to be critical.”
entertainment	(0.020) play, (0.017) watch, (0.016) safe, (0.016) time, (0.015) game, (0.014) read, (0.013) stay, (0.012) book, (0.011) write, (0.010) hope, (0.010) video, (0.010) movie, (0.009) start, (0.008) team, (0.007)	“I'm used to reading multiple books (up to 10) at a time, and I know that sometimes I can't start a book/get into a book when I've got the others to finish, but that's not it w this one. and I'm not having trouble starting new

	season, (0.007) event, (0.007) lot, (0.007) win, (0.007) forward, (0.006) release	books - if anything, this pandemic has made me read.”
death	(0.085) death, (0.030) die, (0.025) rate, (0.020) report, (0.019) news, (0.011) daily, (0.010) total, (0.010) infection, (0.010) record, (0.010) low, (0.009) disabled, (0.009) county, (0.009) gay, (0.008) day, (0.008) flu, (0.008) count, (0.008) panic, (0.008) attack, (0.008) population, (0.007) figure	<p>“Take a look at death rates from flu, death rates before 2020, and covid death rates following infection rates. there is no question that increased death rates follow increased infection rates.”</p> <p>“Me and bestie were gonna do the amazing race (no way that’s coming back thanks to covid) under the stipulation she did all the eating challenges. I’ll probably try in the future”</p>
wear masks	(0.143) mask, (0.097) wear, (0.078) wear_mask, (0.034) social, (0.019) distancing, (0.017) social_distancing, (0.013) hand, (0.011) distance, (0.011) stay, (0.011) mandate, (0.008) store, (0.007) vaccinate, (0.007) public, (0.007) safe, (0.007) protect, (0.007) refuse, (0.006) social_distance, (0.005) wash, (0.005) guideline, (0.004) time	<p>“I don’t know which year this picture is from, but I hope it’s 2019, since no one is wearing a mask or social distancing. we’re still waiting for a picture of you, pete, wearing a mask. be a good example.”</p> <p>“I thought I could go on a camping trip but I can't go to any state that's remotely interesting because you idiots won't wear a mask.”</p>
children and education	(0.074) school, (0.058) kid, (0.035) child, (0.021) student, (0.020) class, (0.017) parent, (0.016) online, (0.015) teacher, (0.010) college, (0.009) cancel, (0.008) teach, (0.008) asian, (0.007) south, (0.006) send, (0.005) rapid, (0.005) university, (0.005) attend, (0.005) excited, (0.005) board, (0.005) court	<p>“Please cancel all state board exam and 12 class board examplease cancel itwe are also student we can be also affected by corona.. why you didn't cancel all state board exam and 12 class board exam ...”</p> <p>“No school for my grandchildren until they have been vaccinated against the coronavirus. home school is going to be the only way for now. ”</p>
politics	(0.048) trump, (0.025) american, (0.019) vote, (0.018) lie, (0.017) biden, (0.015) die, (0.013) republican, (0.012) president, (0.011) kill, (0.010) dead, (0.010) election, (0.010) china, (0.009) call, (0.009) country, (0.008) america, (0.008) response, (0.007) gop, (0.007) blame, (0.007) care, (0.006) hoax	<p>“Just a reminder under trump is out of control with no coordinated national response. over 141,000 americans.”</p> <p>“Disappointed? this administration did nothing preemptive for this pandemic. governed by an old president, the health system on the hands of the very meaning of incompetence. What do we get? hmm”</p>
lockdown	(0.084) lockdown, (0.015) time, (0.014) country, (0.012) travel, (0.011) start, (0.011) month, (0.010) restriction, (0.009) week, (0.009) government, (0.008) lock, (0.008) close, (0.008) rule, (0.007) happen,	“England is no longer a free country. lord sumption described boris johnson's covid restrictions as the "most significant interference with personal freedom in the history of our country". Boris is a tyrant.”

	(0.007) live, (0.007) day, (0.006) plan, (0.006) city, (0.005) outbreak, (0.005) march, (0.005) summer	“There's no point at all. that's what happened here in germany, we called it here light lockdown. only pubs, restaurants and cinemas were closed in november. and look where we are now. we've got 2 weeks straight now the highest infection and mortality rate in eu”
gender and race	(0.013) woman, (0.013) white, (0.012) life, (0.011) protest, (0.010) black, (0.009) police, (0.008) texas, (0.008) florida, (0.008) epidemic, (0.007) mental_health, (0.007) medium, (0.007) kill, (0.006) super, (0.006) anti, (0.006) spread, (0.006) war, (0.005) human, (0.005) crisis, (0.005) aid, (0.005) racism	“Yall, its 2020. we all really, really need to understand how gentrification works, and how gentrification is violent for black communities in atlanta. thanks to disaster capitalism, there is always a land grab post any major crisis and the same is poised to happen with covid.” “there's a fucking pandemic going on, but sure blame the unions cause it's always about trashing teachers who are overwhelmingly women and seen as garbage.”

1 **Table S4.** Topic distribution over different random samples.

	Original sample (%) (n=1,742,385)	Sample A (%) (n=1,742,444)	Sample B (%) (n=3,484,880)	P value*
Topics				0.081
Vaccine	206,950 (11.88)	205,159 (11.77)	412,512 (11.84)	
Friends and family	228,812 (13.13)	229,475 (13.17)	458,420 (13.15)	
Economy	139,718 (8.02)	138,947 (7.97)	278,209 (7.98)	
COVID-19 positive	119,737 (6.87)	119,589 (6.86)	239,316 (6.87)	
Medicine	131,090 (7.52)	130,783 (7.51)	261,970 (7.52)	
Entertainment	95,454 (5.48)	95,876 (5.50)	192,091 (5.51)	
Death	103,626 (5.95)	103,769 (5.96)	207,392 (5.95)	
Wear masks	144,446 (8.30)	144,854 (8.31)	289,289 (8.30)	
Children and education	53,268 (3.06)	53,871 (3.10)	107,920 (3.10)	
Politics	195,601 (11.23)	195,544 (11.22)	391,289 (11.23)	
Lockdown	218,585 (12.55)	218,394 (12.53)	436,592 (12.53)	
Gender and race	105,098 (6.03)	106,183 (6.10)	209,880 (6.02)	

2 *P values were calculated using χ^2 test.

3
4

1 **Table S5.** Lexicon for vaccine incorporation.

Vaccine	Lexicon
AZ	Astra Zeneca AstraZeneca Astrazeneca astrazeneca astra zeneca ASTRAZENECA astra Zeneca Astra Zenica AstraZenica Astra zenica astra zenica Astrazenica astrazenica astraderna Astra vaccination AstraZenna Astragenica AztraSeneca AztraZeneca Aztrazeneca AstroZeneca Astrozeneca AsatraZeneca Astra-Zeneca Astra-Zenica Astra-Zeneca vaccine astrozeneca aztrazeneca astazenica Astera Zeneca GenXZeneca shot Astra vaccine Astra jab AstraZ Astra Z AZ Vax AZ vax AZ vac AZ vaxxed AZ vaccine AZ vaccines AZ Vaccine AZ vaccination AZ vaccine shot AZ vaccinated AZ jab AZ jabs AZ Jabs AZ shot AZ shots AZ covid vaccine AZ Covid jab AZ Covid vaccine AZ COVID shot AZ COVID vax AZ COVID vaccines AZ COVID vaccine AZ covid jab AZ No jab AZ-Oxford vaccine AZ dose AZ vacc AZ one A-Z vaccine Az vaccine az vaccine Astra one AZN vaccine AZN COVID vaccine AZN COVID-19 Vaccine AZN COVID-19 vaccine AZN LN vaccine AZN shot AZN AZD1222 AZD1222 vaccine AZ/Oxford ChAdOx1 nCoV-19 vaccine ChAdOx1 nCoV ChAdOx1 Vaxzevria COVIDSHIELD COVISHIELD COVID shield CoviShield Covid Shield Covidshield Covishield Covid shield covidshield covishield covid shield Oxford Vaccine Oxford vaccine Oxford vaccines Oxford vaccination OXFORD vaccine OxfordAZ vaccine oxford VACCINE Oxford Coronavirus Vaccine Oxford Coronavirus vaccine Oxford coronavirus vaccine Oxford COVID vaccine Oxford COVID-19 vaccine Oxford COVID shot Oxford Covid-19 vaccine Oxford University COVID-19 vaccine Oxford University vaccine Oxford Covid vaccine Oxford covid vaccine OxfordVaccine oxfordvaccine Oxford jab Oxford vax Oxford-AZ Oxford/AZ Oxford AZ
Moderna	MODERNA Moderna Modernas Modera ModeRNA Maderna vaccine Maderna moderna maderna SpikeVax Spikevax
Pfizer	Pfizer pfizer PFIZER Pfyzer Pfeizer Pfizers Pfiser PFizer Pfitzer pfiser Pfizers vaccine Phizer shots phizer shot Pfizer jabbed Pfeizer Pfiizer PHIZER Pfzier Pfzer pfeizer Pzizer pzifer Pizer Pzifer Pifzer fizer Phizener Phizer phizer P-Brand Vaccine BioNTech Biontech BioNtech biotech BioNTTech BNTX vaccine BNT162b2 BNT162 BionTech BNTX biontech Pz/B Tec vaccine PFE vaccine PPE vaccines PFZ vaccine PFE.N PFE pfe COMIRNATY Comirnaty
JNJ	JohnsonAndJohnson JohnsonandJohnson Johnsonandjohnson Johnson and Johnson Johnson & Johnson Johnson&Johnson Johnson Johnson JohnsonJohnson JOHNSON & JOHNSON JOHNSON JOHNSON Johnson Johnson coronavirus vaccine Johnson Johnson's vaccine Johnson Johnson vaccines Johnson Johnson shot Johnson and johnson Johnson Johnson Covid-19 vaccine Johnson Johnson Covid vaccine Johnson Johnson COVID vaccine johnson and johnson johnson and johnson vaccine johnsonandjohnson johnson And Johnson johnson johnson JOHNSON JOHNSON COVID-19 VACCINE JOHNSON

<p>AND JOHNSON Johnson for vaccine Johnson COVID-19 Vaccine Johnson COVID-19 vaccine Johnson's COVID-19 Vaccine Johnson's Covid-19 vaccine Johnson's one-shot Janssen COVID-19 Vaccine Janssen COVID-19 vaccine JOHNSON COVID-19 VACCINE Johnson vaccine Janssen vaccine Janssen janssen JANSSEN Janssen/J J JNJ COVID-19 Vaccine JNJ COVID-19 VACCINE JNJ COVID-19 vaccine JNJ Covid vaccine JNJ Covid-19 vaccine JNJ Vaccine JNJ covid vaccine JNJ shot JNJ single-dose vaccine JNJ vaccine JNJ vaccine shot JNJ vaccine single shot JNJ vax JnJ vaccine jnj vaccine J J/Janssen COVID-19 vaccine J J Covid-19 Vaccine J J Covid-19 vaccine J J COVID - 19 VACCINE J J COVID-19 vaccine J J COVID vaccine J J covid vaccine J J Vax J J jab J J vac J J Vaccine J J vaccine J J vaccines J J VACCINE J J vaccination J J vaccinated J J vaccinations J J vaxx J J vaxxed J J shot J J shots J J one shot J J Covid shot J J covid shot J J Covid vaccine J J's vaccine J J's vaccine J J/Janssen vaccine JJ vax J and J shot J and J J and J vaccine J J vax J/J j j shot J J/Janssen j j vax j j vaccine j and j J&J JandJ J & J vaccine JJ shots JJ shot JJ vaccine J vaccine</p>

1

2

3

4

Table S6. Keywords for the selection of users who self-reported positive. Users who post word combinations from the "Verb" and "Noun" columns or keywords from the "Other" column on Twitter are identified as infected.

Verb	Noun	Other
<p>get got have had diagnose diagnosed diagnose with diagnosed with catch caught by infected by</p>	<p>covid corona ncov covid-19 covid19 coronavirus koronavirus sars-cov-2 covd virus a virus the virus</p>	<p>was positive were positive test positive tested positive identified by test recognized by test</p>

5

6

1 References

- 2 1. Marmot, M. and J. Allen, *COVID-19: exposing and amplifying inequalities*. J Epidemiol
3 Community Health, 2020. **74**(9): p. 681-682.
- 4 2. Gibb, J.K., L.Z. DuBois, S. Williams, L. McKerracher, R.-P. Juster, and J. Fields, *Sexual
5 and gender minority health vulnerabilities during the COVID-19 health crisis*. American
6 Journal of Human Biology, 2020. **32**(5): p. e23499.
- 7 3. Phillips Ii, G., D. Felt, M.M. Ruprecht, X. Wang, J. Xu, E. Pérez-Bill, R.M. Bagnarol, J.
8 Roth, C.W. Curry, and L.B. Beach, *Addressing the disproportionate impacts of the COVID-
9 19 pandemic on sexual and gender minority populations in the United States: actions
10 toward equity*. LGBT health, 2020. **7**(6): p. 279-282.
- 11 4. Beach, L.B., T.A. Elasy, and G. Gonzales, *Prevalence of self-reported diabetes by sexual
12 orientation: Results from the 2014 Behavioral Risk Factor Surveillance System*. Lgbt Health,
13 2018. **5**(2): p. 121-130.
- 14 5. O'Neill, K., *Health vulnerabilities to COVID-19 among LGBT adults in California*. 2020.
- 15 6. Heslin, K.C., *Sexual orientation disparities in risk factors for adverse COVID-19-related
16 outcomes, by race/ethnicity—Behavioral Risk Factor Surveillance System, United States,
17 2017–2019*. MMWR. Morbidity and Mortality Weekly Report, 2021. **70**.
- 18 7. Konnoth, C., *Supporting LGBT communities in the COVID-19 pandemic*. 2020). Assessing
19 Legal Responses to COVID-19. Boston: Public Health Law Watch, U of Colorado Law
20 Legal Studies Research Paper, 2020(20-47).
- 21 8. Anderson, R.M., C. Vegvari, J. Truscott, and B.S. Collyer, *Challenges in creating herd
22 immunity to SARS-CoV-2 infection by mass vaccination*. The Lancet, 2020. **396**(10263): p.
23 1614-1616.
- 24 9. Britton, T., F. Ball, and P. Trapman, *A mathematical model reveals the influence of
25 population heterogeneity on herd immunity to SARS-CoV-2*. science, 2020. **369**(6505): p.
26 846-849.
- 27 10. Garg, I., H. Hanif, N. Javed, R. Abbas, S. Mirza, M.A. Javaid, S. Pal, R. Shekhar, and A.B.
28 Sheikh, *COVID-19 vaccine hesitancy in the LGBTQ+ population: a systematic review*.
29 Infectious Disease Reports, 2021. **13**(4): p. 872-887.
- 30 11. Carson, S.L., A. Casillas, Y. Castellon-Lopez, L.N. Mansfield, J. Barron, E. Ntekume, R.
31 Landovitz, S.D. Vassar, K.C. Norris, and S.M. Dubinett, *COVID-19 vaccine decision-
32 making factors in racial and ethnic minority communities in Los Angeles, California*. JAMA
33 network open, 2021. **4**(9): p. e2127582-e2127582.
- 34 12. Machingaidze, S. and C.S. Wiysonge, *Understanding COVID-19 vaccine hesitancy*. Nature
35 medicine, 2021. **27**(8): p. 1338-1339.
- 36 13. Gonzales, G., E.L. de Mola, K.A. Gavulic, T. McKay, and C. Purcell, *Mental health needs
37 among lesbian, gay, bisexual, and transgender college students during the COVID-19
38 pandemic*. Journal of Adolescent Health, 2020. **67**(5): p. 645-648.
- 39 14. Adamson, T., M. Hanley, S. Baral, C. Beyrer, S. Wallach, and S. Howell, *Rapid,
40 application-based survey to characterise the impacts of COVID-19 on LGBTQ+
41 communities around the world: an observational study*. BMJ open, 2022. **12**(4): p. e041896.
- 42 15. Lynch, K.E., J.C. Shipherd, E. Gatsby, B. Viernes, S.L. DuVall, and J.R. Blosnich, *Sexual
43 orientation-related disparities in health conditions that elevate COVID-19 severity*. Annals
44 of Epidemiology, 2022. **66**: p. 5-12.
- 45 16. Rivera, A.S., M. Plank, A. Davis, M.J. Feinstein, L.K. Rusie, and L.B. Beach, *Assessing
46 widening disparities in HbA1c and systolic blood pressure retesting during the COVID-19
47 pandemic in an LGBTQ+-focused federally qualified health center in Chicago: a
48 retrospective cohort study using electronic health records*. BMJ Open Diabetes Research
49 and Care, 2022. **10**(6): p. e002990.

- 1 17. Wu, J., X. Liu, M. Li, W. Li, Z. Su, S. Lin, L. Garay, Z. Zhang, Y. Zhang, and Q. Zeng, *Clinical text datasets for medical artificial intelligence and large language models—a systematic review*. *NEJM AI*, 2024; p. AIra2400012.
- 2 18. Rains, S.A., *Big data, computational social science, and health communication: a review and agenda for advancing theory*. *Health communication*, 2020. **35**(1): p. 26-34.
- 3 19. Jordan, S.E., S.E. Hovet, I.C.-H. Fung, H. Liang, K.-W. Fu, and Z.T.H. Tse, *Using Twitter for public health surveillance from monitoring and prediction to public response*. *Data*, 2018. **4**(1): p. 6.
- 4 20. Wu, J., X. Wu, Y. Hua, S. Lin, Y. Zheng, and J. Yang. *Exploring social media for early detection of depression in covid-19 patients*. in *Proceedings of the ACM Web Conference 2023*. 2023.
- 5 21. Tsao, S.-F., H. Chen, T. Tisseverasinghe, Y. Yang, L. Li, and Z.A. Butt, *What social media told us in the time of COVID-19: a scoping review*. *The Lancet Digital Health*, 2021. **3**(3): p. e175-e194.
- 6 22. Zhang, Y., H. Lyu, Y. Liu, X. Zhang, Y. Wang, and J. Luo, *Monitoring depression trends on Twitter during the COVID-19 pandemic: observational study*. *JMIR infodemiology*, 2021. **1**(1): p. e26769.
- 7 23. Crocamo, C., M. Viviani, L. Famiglini, F. Bartoli, G. Pasi, and G. Carrà, *Surveilling COVID-19 emotional contagion on twitter by sentiment analysis*. *European Psychiatry*, 2021. **64**(1): p. e17.
- 8 24. Sinnenberg, L., A.M. Bittenheim, K. Padrez, C. Mancheno, L. Ungar, and R.M. Merchant, *Twitter as a tool for health research: a systematic review*. *American journal of public health*, 2017. **107**(1): p. e1-e8.
- 9 25. Shah, Z., D. Surian, A. Dyda, E. Coiera, K.D. Mandl, and A.G. Dunn, *Automatically appraising the credibility of vaccine-related web pages shared on social media: a Twitter surveillance study*. *Journal of medical Internet research*, 2019. **21**(11): p. e14007.
- 10 26. Hua, Y., H. Jiang, S. Lin, J. Yang, J.M. Plasek, D.W. Bates, and L. Zhou, *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. **29**(10): p. 1668-1678.
- 11 27. Hamamsy, T. and R. Bonneau, *Twitter activity about treatments during the COVID-19 pandemic: case studies of remdesivir, hydroxychloroquine, and convalescent plasma*. *medRxiv*, 2020; p. 2020.06. 18.20134668.
- 12 28. Chew, C. and G. Eysenbach, *Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak*. *PloS one*, 2010. **5**(11): p. e14118.
- 13 29. Masri, S., J. Jia, C. Li, G. Zhou, M.-C. Lee, G. Yan, and J. Wu, *Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic*. *BMC public health*, 2019. **19**: p. 1-14.
- 14 30. Müller, M., M. Salathé, and P.E. Kummervold, *Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter*. *Frontiers in artificial intelligence*, 2023. **6**: p. 1023281.
- 15 31. Jiang, H., Y. Hua, D. Beeferman, and D. Roy, *Annotating the Tweebank corpus on named entity recognition and building NLP models for social media analysis*. *arXiv preprint arXiv:2201.07281*, 2022.
- 16 32. Zhou, P., Z. Wang, D. Chong, Z. Guo, Y. Hua, Z. Su, Z. Teng, J. Wu, and J. Yang, *Mets-cov: A dataset of medical entity and targeted sentiment on covid-19 related tweets*. *Advances in Neural Information Processing Systems*, 2022. **35**: p. 21916-21932.
- 17 33. Li, W., Y. Hua, P. Zhou, Z. Li, X. Xu, and J. Yang, *Characterizing Public Sentiments and Drug Interactions during COVID-19: A Pretrained Language Model and Network Analysis of Social Media Discourse*. *medRxiv*, 2024; p. 2024.06. 06.24308537.

- 1 34. Chen, E., K. Lerman, and E. Ferrara, *Tracking social media discourse about the covid-19*
2 *pandemic: Development of a public coronavirus twitter data set*. JMIR public health and
3 surveillance, 2020. **6**(2): p. e19273.
- 4 35. Mayer, K.H., J.B. Bradford, H.J. Makadon, R. Stall, H. Goldhammer, and S. Landers,
5 *Sexual and gender minority health: what we know and what needs to be done*. American
6 journal of public health, 2008. **98**(6): p. 989-995.
- 7 36. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation*. Journal of machine
8 Learning research, 2003. **3**(Jan): p. 993-1022.
- 9 37. Řehůřek, R. and P. Sojka, *Software framework for topic modelling with large corpora*. 2010.
- 10 38. Sievert, C. and K. Shirley. *LDavis: A method for visualizing and interpreting topics*. in
11 *Proceedings of the workshop on interactive language learning, visualization, and interfaces*.
12 2014.
- 13 39. Kessler, J.S., *Scattertext: a browser-based tool for visualizing how corpora differ*. arXiv
14 preprint arXiv:1703.00565, 2017.
- 15 40. Chen, C., Z. Teng, and Y. Zhang. *Inducing target-specific latent structures for aspect*
16 *sentiment classification*. in *Proceedings of the 2020 conference on empirical methods in*
17 *natural language processing (EMNLP)*. 2020.
- 18 41. Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep*
19 *bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805,
20 2018.
- 21 42. Yang, J. and Y. Zhang, *NCRF++: An open-source neural sequence labeling toolkit*. In
22 *Proceedings of ACL 2018, System Demonstrations* (pp. 74-79).
- 23 43. Wu, J., L. Wang, Y. Hua, M. Li, L. Zhou, D.W. Bates, and J. Yang, *Trend and co-*
24 *occurrence network of COVID-19 symptoms from large-scale social media data:*
25 *Infoveillance study*. Journal of Medical Internet Research, 2023. **25**: p. e45419.
- 26 44. Li, M., Y. Hua, Y. Liao, L. Zhou, X. Li, L. Wang, and J. Yang, *Tracking the impact of*
27 *covid-19 and lockdown policies on public mental health using social media: Infoveillance*
28 *study*. Journal of Medical Internet Research, 2022. **24**(10): p. e39676.
- 29 45. Thelwall, M. and S. Thelwall, *Twitter during COVID-19: George Floyd opening a space to*
30 *address systematic and institutionalized racism?* Available at SSRN 3764867, 2021.
- 31 46. Reny, T.T. and B.J. Newman, *The opinion-mobilizing effect of social protest against police*
32 *violence: Evidence from the 2020 George Floyd protests*. American political science review,
33 2021. **115**(4): p. 1499-1507.
- 34 47. Baccini, L., A. Brodeur, and S. Weymouth, *The COVID-19 pandemic and the 2020 US*
35 *presidential election*. Journal of population economics, 2021. **34**: p. 739-767.
- 36 48. Sears, B., K.J. Conron, and A.R. Flores, *The impact of the fall 2020 COVID-19 surge on*
37 *LGBT adults in the US*. 2021.
- 38 49. Tahir, A., L. Cheng, P. Sheth, and H. Liu, *Improving vaccine stance detection by combining*
39 *online and offline data*. arXiv preprint arXiv:2208.04491, 2022.
- 40 50. Lyu, H., J. Wang, W. Wu, V. Duong, X. Zhang, T.D. Dye, and J. Luo, *Social media study*
41 *of public opinions on potential COVID-19 vaccines: informing dissent, disparities, and*
42 *dissemination*. Intelligent medicine, 2022. **2**(01): p. 1-12.
- 43 51. Kuehn, B.M., *Racial and Ethnic, Gender Disparities Seen in LGBT COVID-19 Vaccination*
44 *Rates*. JAMA, 2022. **327**(10): p. 910-910.
- 45 52. McNaghten, A., *COVID-19 vaccination coverage and vaccine confidence by sexual*
46 *orientation and gender identity—United States, August 29–October 30, 2021*. MMWR.
47 Morbidity and Mortality Weekly Report, 2022. **71**.
- 48 53. Mallory, C., B. Sears, and A. Flores, *COVID-19 and LGBT Adults Ages 45 and Older in the*
49 *US*. 2021.

- 1 54. Teixeira da Silva, D., K. Biello, W.Y. Lin, P.K. Valente, K.H. Mayer, L. Hightow-Weidman,
2 and J.A. Bauermeister, *COVID-19 vaccine acceptance among an online sample of sexual*
3 *and gender minority men and transgender women*. *Vaccines*, 2021. **9**(3): p. 204.
- 4 55. Jaiswal, J., K.D. Krause, R.J. Martino, P.A. D'Avanzo, M. Griffin, C.B. Stults, A.G. Karr,
5 and P.N. Halkitis, *SARS-CoV-2 vaccination hesitancy and behaviors in a national sample*
6 *of people living with HIV*. *AIDS patient care and STDs*, 2021. **36**(1): p. 34-44.
- 7 56. Aiyegbusi, O.L., S.E. Hughes, G. Turner, S.C. Rivera, C. McMullan, J.S. Chandan, S.
8 Haroon, G. Price, E.H. Davies, and K. Nirantharakumar, *Symptoms, complications and*
9 *management of long COVID: a review*. *Journal of the Royal Society of Medicine*, 2021.
10 **114**(9): p. 428-442.
- 11 57. Han, Q., B. Zheng, L. Daines, and A. Sheikh, *Long-term sequelae of COVID-19: a*
12 *systematic review and meta-analysis of one-year follow-up studies on post-COVID*
13 *symptoms*. *Pathogens*, 2022. **11**(2): p. 269.
- 14 58. Huang, Y., M.D. Pinto, J.L. Borelli, M. Asgari Mehrabadi, H.L. Abraham, N. Dutt, N.
15 Lambert, E.L. Nurmi, R. Chakraborty, and A.M. Rahmani, *COVID symptoms, symptom*
16 *clusters, and predictors for becoming a long-hauler looking for clarity in the haze of the*
17 *pandemic*. *Clinical nursing research*, 2022. **31**(8): p. 1390-1398.
- 18 59. Zhang, H., C. Zang, Z. Xu, Y. Zhang, J. Xu, J. Bian, D. Morozuyuk, D. Khullar, Y. Zhang,
19 and A.S. Nordvig, *Data-driven identification of post-acute SARS-CoV-2 infection*
20 *subphenotypes*. *Nature Medicine*, 2023. **29**(1): p. 226-235.
- 21 60. Sahin, T., A. Ayyildiz, K. Gencer-Atalay, C. Akgün, H.M. Özdemir, and B. Kuran, *Pain*
22 *symptoms in COVID-19*. *American journal of physical medicine & rehabilitation*, 2021.
23 **100**(4): p. 307-312.
- 24 61. Weng, L.-M., X. Su, and X.-Q. Wang, *Pain symptoms in patients with coronavirus disease*
25 *(COVID-19): a literature review*. *Journal of Pain Research*, 2021: p. 147-159.
- 26 62. Wood, C.I., Z. Yu, D.-A. Sealy, I. Moss, E. Zigbuo-Wenzler, C. McFadden, D. Landi, and
27 A.M. Brace, *Mental health impacts of the COVID-19 pandemic on college students*. *Journal*
28 *of American college health*, 2024. **72**(2): p. 463-468.
- 29 63. Li, Y., Q. Qin, Q. Sun, L.D. Sanford, A.N. Vgontzas, and X. Tang, *Insomnia and*
30 *psychological reactions during the COVID-19 outbreak in China*. *Journal of Clinical Sleep*
31 *Medicine*, 2020. **16**(8): p. 1417-1418.
- 32 64. Voitsidis, P., I. Gliatas, V. Bairachtari, K. Papadopoulou, G. Papageorgiou, E. Parlapani, M.
33 Syngelakis, V. Holeva, and I. Diakogiannis, *Insomnia during the COVID-19 pandemic in a*
34 *Greek population*. *Psychiatry research*, 2020. **289**: p. 113076.
- 35 65. Morin, C.M., L.-A. Vézina-Im, H. Ivers, J.-A. Micoulaud-Franchi, P. Philip, M. Lamy, and
36 J. Savard, *Prevalent, incident, and persistent insomnia in a population-based cohort tested*
37 *before (2018) and during the first-wave of COVID-19 pandemic (2020)*. *Sleep*, 2022. **45**(1):
38 p. zsab258.
- 39 66. Salerno, J.P., N.D. Williams, and K.A. Gattamorta, *LGBTQ populations: Psychologically*
40 *vulnerable communities in the COVID-19 pandemic*. *Psychological trauma: Theory,*
41 *research, practice, and policy*, 2020. **12**(S1): p. S239.
- 42 67. Zhai, Y. and X. Du, *Disparities and intersectionality in social support networks: addressing*
43 *social inequalities during the COVID-19 pandemic and beyond*. *Humanities and Social*
44 *Sciences Communications*, 2022. **9**(1): p. 1-5.
- 45 68. Mislove, A., S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist. *Understanding the*
46 *demographics of Twitter users*. in *Proceedings of the international AAAI conference on web*
47 *and social media*. 2011.
- 48 69. Hua, Y., J. Wu, S. Lin, M. Li, Y. Zhang, D. Foer, S. Wang, P. Zhou, J. Yang, and L. Zhou,
49 *Streamlining social media information extraction for public health research with deep*
50 *learning*. *Journal of the American Medical Informatics Association*, 2024: p. ocae118.