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

Authors
Zhiyun Zhang\textsuperscript{1}, Yining Hua\textsuperscript{2,3,4}, Peilin Zhou\textsuperscript{5}, Shixu Lin\textsuperscript{1}, Minghui Li\textsuperscript{1}, Yujie Zhang\textsuperscript{1}, Li Zhou\textsuperscript{4}, Yanhui Liao\textsuperscript{6}, Jie Yang\textsuperscript{1*}

Affiliations
\textsuperscript{1} Department of Big Data in Health Science School of Public Health, Zhejiang University School of Medicine, Hangzhou, China.
\textsuperscript{2} Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, USA.
\textsuperscript{3} Department of Biomedical Informatics, Harvard Medical School, Boston, USA.
\textsuperscript{4} Division of General Internal Medicine and Primary Care, Department of Medicine, Brigham and Women’s Hospital, Boston, USA.
\textsuperscript{5} Thrust of Data Science and Analytics, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China.
\textsuperscript{6} Department of Psychiatry, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China.

Corresponding author
Jie Yang, Ph.D.
School of Public Health, Zhejiang University School of Medicine, Hangzhou, China
Address: 866 Yuhangtang Rd, Hangzhou, P.R. China 310058
Email: jieynlp@gmail.com
Abstract

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

MAIN TEXT

1. Introduction

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

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

However, research on the health of SGD individuals during the COVID-19 pandemic faces a notable gap in both depth and breadth. Most studies rely on online surveys and questionnaires [13, 14], constrained by the inherent biases of questionnaire design and the prolonged intervals of data collection. Moreover, research utilizing electronic health records typically focuses on specific symptoms of severe COVID-19, influenced by hospital admission rates and delays in gathering SGD data [15, 16], thus lacking in generalizability. Additionally, the access to electronic health records is highly restricted, limiting the coverage of patients [17]. A major challenge in conducting comprehensive health status
analysis for SGD populations lies in selecting large-scale, representative cohorts, which underscores the potential of alternative data sources, such as social media, which has been increasingly validated as a valuable tool in public health research [18-20]. Social media-related studies cover topics ranging from mental healthcare [21-23] to disease symptoms [24] and public acceptance of treatments [25-27], both in the context of COVID-19 and previous health crises such as H1N1 [28] and Zika [29]. Meanwhile, new deep learning-based language models, pipelines, and datasets [30-33] offer opportunities to analyze the massive textual information from social media platforms. This synergy between natural language processing (NLP) and social media analytics opens up novel avenues for research that span both data collection and analytical interpretation [18].

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

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

2. Methods

Experimental Design

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

Data collection and selection

This study collected tweets through leveraging tweet IDs from a public coronavirus Twitter dataset [34], which follows specified accounts and collects real-time tweets mentioning specific keywords. We instituted a filtering process where tweets containing URLs were excluded to attenuate the impact of news and automated bot activities. Subsequently, we focused on identifying tweets from SGD users, including lesbian, gay, bisexual, transgender, queer, intersex, and asexual individuals [35]. SGD users were filtered through user profiles using keyword filtering and regular expression matching (Table S1): 1) User profiles must contain SGD-related keywords. 2) There should be no negation words before or after the keywords. 3) The keywords should not be preceded or followed by terms such as "advocator" and "supporter" as some users may advocate for SGD rights without necessarily
being SGD themselves. A manual validation process was conducted on a subset of 500 selected SGD users, achieving a classification accuracy of 93.8%. We also evaluated the baseline characteristics of geographic information on the validation subset (Table S2).

**Statistical Analysis**

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

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

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

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

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

Tweets posted before and after 30 days of their self-report date were collected for users who self-reported COVID-19 diagnoses. We then screened these tweets using a COVID-19
symptom lexicon developed by Wu et al. [43], which contains commonly used synonyms and colloquial variants on social media that pertain to symptoms and their associated affected organs or systems. For the identification of mental health-related tweets, we utilized an exhaustive mental health lexicon [44] which has been rigorously validated by professionals in the fields of psychiatry and psychology. This lexicon encompasses 231 keywords distributed across four major mental health conditions: anxiety, depression, insomnia, and substance use disorders.

3. Results

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

Topic distribution and discrepancies

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

When comparing topic frequencies between SGD and non-SGD users using χ² test with Bonferroni adjusted significance level P < 0.004, the data indicate that discussions about "friends and family" were significantly more prevalent among the former (20.5% vs 13.1%, P<0.001) (Figure 1B). For a deeper insight into the discussion variations, we used Scattertext [39] to visualize the word frequencies (Figure S3). The result highlights that terms regarding family members occur frequently, with SGD users often mentioning "partner" while non-SGD users more frequently use terms like "daughter" and "baby". Furthermore, SGD users express a range of emotions more frequently, especially negative ones like "anxiety", "upset", "angry", "depression". For other topics, SGD users were more likely to talk about "wear masks" (10.1% vs 8.3%, P<0.001) and "COVID-19 positive" (8.6% vs 6.9%, P<0.001), while non-SGD users discussed other topics like "vaccine" (9.6% vs 11.9%, P<0.001) and "lockdown" (10.3% vs 12.5%, P<0.001) more often.
Fig. 1. Topic distributions in COVID-19-related tweets. (A) The topic proportion distribution of SGD user-posted tweets over time. (B) Topic distributions of tweets posted by SGD users and non-SGD users.

Attitude Towards COVID-19 Vaccines

Table 1 shows the distribution of three-category sentiments (positive, neutral, and negative), wherein the majority were characterized as neutral. Notably, sentiments toward the Pfizer vaccine exhibited the highest frequency of positive evaluations. Contrarily, attitudes toward the AstraZeneca vaccine appeared the most polarized among SGD users when contrasted with non-SGD users. Utilizing an Independent Samples t-test for statistical analysis, we found that SGD users displayed significantly higher proportions of positive sentiments for 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 < 7.25×10^-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.

<table>
<thead>
<tr>
<th>Vaccine</th>
<th>Positive Tweets, No. (%) [95% CI]</th>
<th>P value^b (SGD greater)</th>
<th>Negative Tweets, No. (%) [95% CI</th>
<th>P value^b (SGD smaller)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SGD</td>
<td>Non-SGD</td>
<td>SGD</td>
<td>Non-SGD</td>
</tr>
<tr>
<td>Moderna</td>
<td>610 (12.6) [11.6-13.5]</td>
<td>28828 (8.6) [8.5-8.7]</td>
<td>&lt;0.001</td>
<td>182 (3.7) [3.2-4.3]</td>
</tr>
<tr>
<td>Pfizer</td>
<td>984 (12.9) [12.2-13.7]</td>
<td>58109 (8.8) [8.7-8.8]</td>
<td>&lt;0.001</td>
<td>327 (4.3) [3.8-4.8]</td>
</tr>
<tr>
<td>AstraZeneca</td>
<td>158 (11.0) [9.4-12.7]</td>
<td>16227 (3.8) [3.8-3.9]</td>
<td>&lt;0.001</td>
<td>68 (4.8) [3.7-5.9]</td>
</tr>
<tr>
<td>Johnson &amp; John</td>
<td>208 (9.3) [8.1-10.5]</td>
<td>11482 (6.5) [6.4-6.6]</td>
<td>&lt;0.001</td>
<td>127 (5.6) [4.7-6.6]</td>
</tr>
</tbody>
</table>

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

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

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

<table>
<thead>
<tr>
<th>Mental symptoms</th>
<th>SGD (% tweets), median(IQR)</th>
<th>Non-SGD (% tweets), median (IQR)</th>
<th>W</th>
<th>P-value^b</th>
</tr>
</thead>
<tbody>
<tr>
<td>anxiety</td>
<td>1.585 (0.530)</td>
<td>1.047 (0.158)</td>
<td>317574</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>depression</td>
<td>3.634 (0.885)</td>
<td>3.024 (0.315)</td>
<td>302044</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Fig. 3. The proportion of tweets concerning mental health to the total number of tweets. We used a 7-day moving average to smooth the curve.
We used the IQR and Wilcoxon matched-pairs signed-ranks test in the statistical analysis. 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 modeling 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
individuals are more likely to experience more severe symptoms after COVID-19 infection. This may be due to inadequate health insurance coverage and higher-than-average rates of underlying diseases such as diabetes, and asthma, which can increase the risk of severe symptoms [5, 6]. Besides, SGD users mentioned musculoskeletal symptoms (body pain, myalgia pain, arthralgia pain, etc.) at a particularly higher rate compared to non-SGD users. These symptoms are often associated with severe disease as they can be triggered by increased inflammatory factors (e.g., interleukin-6) during infection [60, 61].

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

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

Funding: None


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
Methods

Sensitivity analysis of under-sampled non-SGD tweets in the topic model. To assess the impact of different random seeds on topic distribution, we created two new samples of non-SGD tweets: Sample A, with the same number as the original sample (n = 2,296,289), and Sample B, with twice that number (n = 4,592,578). After preprocessing, 1,742,444 tweets remained in Sample A and 3,484,880 in Sample B. Then we evaluated the distribution of topics in these new corpora using the trained LDA model and compared it with the old distribution using chi square test (Table S4).
Figure S1. Data collection and distribution.

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.
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.
Figure S4. Self-reported COVID-19 positive users filtering process.
**Table S1.** Lexicon and rules for SGD selection. We used regular expressions to remove tweets with stop words before and after the SGD-related keywords to filter out negative expressions and SGD supporters but non-SGD individuals. "|" is used as a delimiter between lexicons.

<table>
<thead>
<tr>
<th>Stop words in front</th>
<th>SGD lexicon</th>
<th>Stop words behind</th>
</tr>
</thead>
<tbody>
<tr>
<td>not</td>
<td>no</td>
<td>none</td>
</tr>
<tr>
<td>lgbt</td>
<td>lesbian</td>
<td>gay</td>
</tr>
<tr>
<td>anti</td>
<td>ally</td>
<td>supporter</td>
</tr>
</tbody>
</table>

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

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

*P values were calculated using χ² test.

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

<table>
<thead>
<tr>
<th>Topics</th>
<th>(value) Keywords</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>vaccine</td>
<td>(0.081) vaccine, (0.022) vaccinate, (0.017) spread, (0.016) flu, (0.013) risk, (0.012) vaccination, (0.012) die, (0.010) variant, (0.010) disease, (0.009) vaccine d, (0.009) prevent, (0.009) booster, (0.008) shoot, (0.008) shot, (0.007) catch, (0.007) science, (0.006) immune, (0.006) chance, (0.006) vaccine</td>
<td>“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.”</td>
</tr>
<tr>
<td>friends and family</td>
<td>(0.026) feel, (0.022) life, (0.020) friend, (0.020) time, (0.018) love, (0.016) family, (0.014) bad, (0.014) stay, (0.014) shit, (0.012) day, (0.011) live, (0.011) die, (0.010) hope, (0.010) lose, (0.008) mom, (0.008) dad</td>
<td>“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.”</td>
</tr>
</tbody>
</table>
0.08) lot, (0.008) start, (0.007) hard, (0.007) god, (0.007) month

---

**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)

---

**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

---

**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

---

**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

---

**“I have no friends and haven’t spoken to anyone in months coz I’d covid and I listen to your baking videos in the background to fill the void, thanks.”**

---

**“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 govt must be given money to augment their lost revenue.”**

---

**“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.”**

---

**“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%. Vaccinations on efforts continue to be critical.”**

---

**“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**
<table>
<thead>
<tr>
<th>Topic</th>
<th>Keywords</th>
<th>Natural Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Season</td>
<td>event, lot, forward, win</td>
<td>“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.”</td>
</tr>
<tr>
<td>Books</td>
<td>- if anything, this pandemic has made me read.</td>
<td></td>
</tr>
<tr>
<td>Lockdown</td>
<td>lockdown, time, rule, happen</td>
<td>“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.”</td>
</tr>
<tr>
<td>Politics</td>
<td>trump, american, president, vote, lie, den, die, republic, president, kill, dead, election, china, call, country, america, response, gop, blame, care, hoax</td>
<td>“Just a reminder under Trump is out of control with no coordinated national response over 141,000 Americans.” “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”</td>
</tr>
<tr>
<td>Children</td>
<td>school, kid, student, class, parent, online, teacher, college, cancel, teach, asian, south, send, rapid, university, attend, excited, board, court</td>
<td>“Please cancel all state board exam and 12 class board exam ... please cancel it ... we are also student we can be also affected by corona. why you didn't cancel all state board exam and 12 class board exam ...” “No school for my grandchildren until they have been vaccinated against the coronavirus. home school is going to be the only way for now.”</td>
</tr>
<tr>
<td>Education</td>
<td>- if anything, this pandemic has made me read.</td>
<td></td>
</tr>
<tr>
<td>Wear masks</td>
<td>mask, wear, distancing, social, distancing, hand, stay, mandate, store, vaccine, public, safe, protect, refuse, social_distance, wash, time</td>
<td>“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.” “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.”</td>
</tr>
<tr>
<td>Children</td>
<td>school, kid, student, class, parent, online, teacher, college, cancel, teach, asian, south, send, rapid, university, attend, excited, board, court</td>
<td>“Please cancel all state board exam and 12 class board exam ... please cancel it ... we are also student we can be also affected by corona. why you didn't cancel all state board exam and 12 class board exam ...” “No school for my grandchildren until they have been vaccinated against the coronavirus. home school is going to be the only way for now.”</td>
</tr>
<tr>
<td>Books</td>
<td>- if anything, this pandemic has made me read.</td>
<td></td>
</tr>
</tbody>
</table>
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

Y'all, 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.”

Table S4. Topic distribution over different random samples.

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

*P values were calculated using χ² test.
<table>
<thead>
<tr>
<th>Vaccine</th>
<th>Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moderna</td>
<td>MODERNA</td>
</tr>
<tr>
<td>Pfizer</td>
<td>Pfizer</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Verb</th>
<th>Noun</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>get</td>
<td>got</td>
<td>have</td>
</tr>
</tbody>
</table>

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.
References


