An Online Tool for Monitoring and Understanding COVID-19 Based on Self-reporting Tweets and Large Language Models

Jiacheng Xie, Ziyang Zhang, Shuai Zeng, Joel Hilliard, Guanghui An, Xiaoting Tang, Lei Jiang, Yang Yu, Xiu-Feng Wan, Dong Xu

Submitted to: Journal of Medical Internet Research
on: June 12, 2024

Table of Contents

An Online Tool for Monitoring and Understanding COVID-19 Based on Self-reporting Tweets and Large Language Models

Jiacheng Xie[1] PhD; Ziyang Zhang[1] MS; Shuai Zeng[1] PhD; Joel Hilliard[1] BA; Guanghui An[2] MD, PhD; Xiaoting Tang[3] MD, PhD; Lei Jiang[1] PhD; Yang Yu[1] PhD; Xiu-Feng Wan[4] MD, PhD; Dong Xu[5] AAAS, PhD

[1]Department of Electrical Engineering and Computer Science University of Missouri Columbia US
[2]School of Acupuncture and Tuina Shanghai University of Traditional Chinese Medicine Shanghai CN
[3]Community Health Service Center Shanghai Pudong New Area Shanghai CN
[4]Department of Molecular Microbiology and Immunology University of Missouri Columbia CN
[5]Department of Electrical Engineering and Computer Science Columbia US

Corresponding Author:
Dong Xu AAAS, PhD
Department of Electrical Engineering and Computer Science
University of Missouri-Columbia
Columbia
US

Abstract

Background: We built a publicly available database of COVID-19-related tweets and extracted information about symptoms and recovery cycles from self-reported tweets. We presented the results of our analysis of infection, reinfection, recovery, and long-term effects of COVID-19 on a weekly- refreshing visualization website.

Objective: We built a publicly available database of COVID-19-related tweets and extracted information about symptoms and recovery cycles from self-reported tweets. We presented the results of our analysis of infection, reinfection, recovery, and long-term effects of COVID-19 on a weekly- refreshing visualization website.

Methods: We used X (formerly Twitter) to collect COVID-related data, from which 9 native-English-speaking annotators annotated a training dataset of COVID-positive self-reporters. We then used large language models to identify positive self-reporters from other unannotated tweets. We employed the Hibert transform to calculate the lead of the prediction curve ahead of the reported curve. Finally, we presented our findings on symptoms, recovery, reinfections, and long-term effects of COVID-19 on the website Covlab.

Results: We collected 9.8 million tweets related to COVID-19 between January 1, 2020, and April 1, 2024, including 469,491 self-reported cases. The predicted number of infection cases by our model is 7.63 days ahead of the official report. In addition to common symptoms, we identified some symptoms that were not included in the list from the Centers for Disease Control and Prevention, such as lethargy and hallucinations. Repeat infections were commonly occurring, with rates of second and third infections at 7.49% and 1.37%, respectively, whereas 0.45% also reported that they had been infected more than 5 times. The average time to recovery has decreased over the years.

Conclusions: Although with some biases and limitations, self-reported tweet data serve as a valuable complement to clinical data, especially in the post-pandemic era dominated by mild cases. Our online analytic platform can play a significant role in continuously tracking COVID-19, finding new uncommon symptoms, detecting and monitoring the manifestation of long-term effects, and providing necessary insights to the public and decision-makers.

(JMIR Preprints 12/06/2024:63190)
DOI: https://doi.org/10.2196/preprints.63190

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

An Online Tool for Monitoring and Understanding COVID-19 Based on Self-reporting Tweets and Large Language Models

Jiacheng Xie,[1,2] Ziyang Zhang,[1,2] Shuai Zeng,[1,2] Joel Hilliard,[1] Guanghui An,[2,3] Xiaoting Tang,[2,4] Lei Jiang,[1,2] Yang Yu,[1,2] Xiu-Feng Wan,[1,2,5,6] Dong Xu,[1,2,*]

[1] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA; [2] Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA; [3] School of Acupuncture and Tuina, Shanghai University of Traditional Chinese Medicine, Shanghai, China; [4] Community Health Service Center Shanghai Pudong New Area, Shanghai, China; [5] NextGen Center for Influenza and Emerging Infectious Diseases, University of Missouri, Columbia, MO, USA; [6] Department of Molecular Microbiology and Immunology, University of Missouri, Columbia, MO, USA
*Corresponding authors

Abstract

Background

Emergence of new SARS-CoV-2 variants, the resulting reinfections, and long COVID continue to impact many people's lives. Tracking websites like the one at Johns Hopkins University no longer report the daily confirmed cases, posing challenges to accurately determining the true extent of infection cases. Many COVID-19 cases with mild symptoms are self-assessed at home and reported on social media, which provides an opportunity to monitor and understand the progression and evolving trends of the disease.

Objectives

We built a publicly available database of COVID-19-related tweets and extracted information about symptoms and recovery cycles from self-reported tweets. We presented the results of our analysis of infection, reinfection, recovery, and long-term effects of COVID-19 on a weekly- refreshing visualization website.

Methods

We used X (formerly Twitter) to collect COVID-related data, from which 9 native-English-speaking annotators annotated a training dataset of COVID-positive self-reporters. We then used large language models to identify positive self-reporters from other unannotated tweets. We employed the Hibert transform to calculate the lead of the prediction curve ahead of the reported curve. Finally, we presented our findings on symptoms, recovery, reinfections, and long-term effects of COVID-19 on the website Covlab.

Results

We collected 7.3 million tweets related to COVID-19 between January 1, 2020, and April 1, 2024, including 469,491 self-reported cases. The predicted number of infection cases by our model is 7.63 days ahead of the official report. In addition to common symptoms, we identified some symptoms that were not included in the list from the Centers for Disease Control and Prevention, such as lethargy and hallucinations. Repeat infections were commonly occurring, with rates of second and third infections at 7.49% and 1.37%, respectively, whereas 0.45% also reported that they had been infected more than 5 times. The average time to recovery has decreased over the years.

Conclusions

Although with some biases and limitations, self-reported tweet data serve as a valuable complement to clinical data, especially in the post-pandemic era dominated by mild cases. Our online analytic platform can play a significant role in continuously tracking COVID-19, finding new uncommon symptoms, detecting and monitoring the manifestation of long-term effects, and providing necessary insights to the public and decision-makers.

Keywords: COVID-19; self-reporting tweets; large language model; Twitter; social media

Introduction

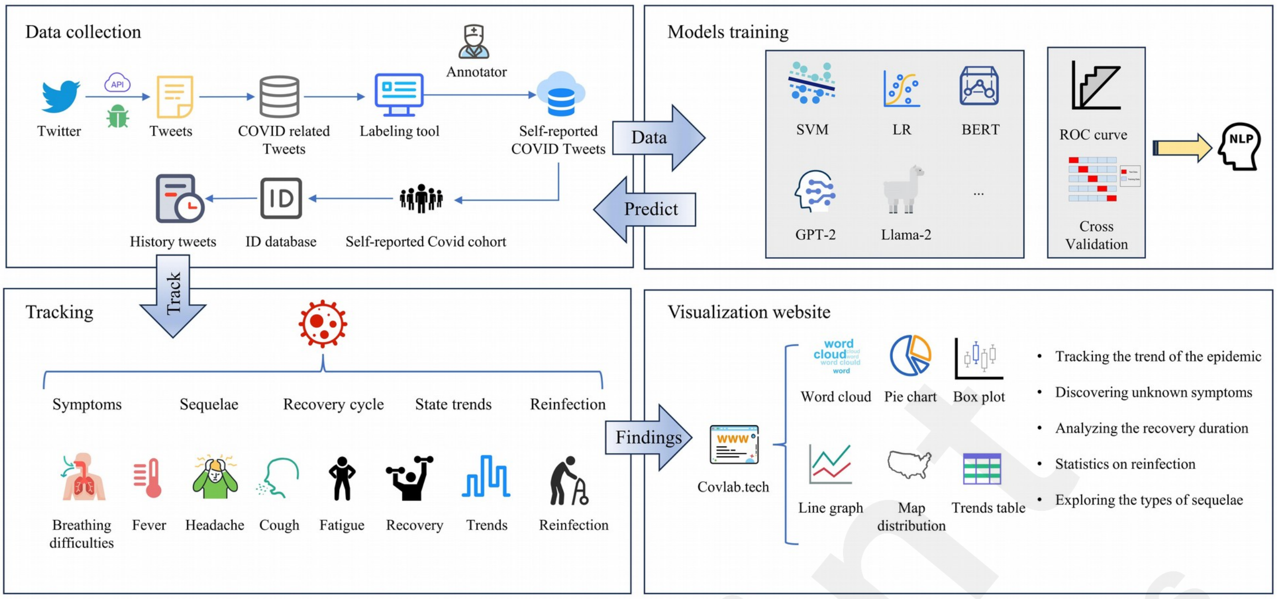
COVID-19 is one of the most severe infectious diseases in human history. Although the World Health Organization (WHO) downgraded the COVID-19 pandemic, declaring it is no longer a global emergency on May 5, 2023 [1], the threat of infection and death remains. As of August 1, 2023, there have been more than 300,000 confirmed cases weekly worldwide, resulting in over 1,000 deaths per week; however, major information-publishing platforms such as that at Johns Hopkins stopped collecting and tracking COVID-19 data worldwide on March 10, 2023 [2]. Therefore, it has become more challenging to identify the actual number and trends of COVID-19 infections daily. New tools are needed to track and report timely awareness and detection of transmission trends and manifestations of COVID-19.

To supplement the shortage of clinical data and gain further insights into the development trends and variant tendencies of COVID-19, researchers have turned to social media, specifically Twitter (X). Early studies [3–7] primarily focused on constructing COVID-19-related tweet databases. However, these works do not provide in-depth analysis of self-reported tweets, and the databases tended to be collected over a short time frame, typically several months, and cannot automatically update. Later, some research endeavors [8–11] shifted their focus toward studying COVID-19 symptoms based on tweets and reported the distribution of symptoms in tweets. However, these studies included limited numbers of collected self-reported cases. Some studies [12–16,7] used tweets to study geographic distribution but did not provide a corresponding time-series analysis or predict the spread of COVID-19. As for the visualization tools for COVID-19, some researchers [17–20] have developed platforms or dashboards to study the trends of COVID-19, but most were based on clinical data. A few tweet-based platforms [18,20,21] showed limited information and failed to provide timely updates.

Reinfection often refers to the phenomenon in which an individual who has recovered from COVID-19 is infected again with the virus [22]. Some researchers [23,24] considered that reinfection is identified when an individual tests positive again through polymerase chain reaction (PCR) testing after a minimum of 90 days of a negative result. However, some studies also suggest this duration should be 30 days [25,26]. Other studies [25,27,28] considered reinfection to be a new positive PCR following 2 consecutive negative PCR tests taken after primary infection. Moreover, the periodicity of reinfection, reinfection rate, and the maximum number of infections are uncertain. Concerning the long-term effects and post-illness symptoms of COVID-19, patient records in clinical data and the relevant research remain limited [29].

In this study, we developed a tool that utilizes self-reported tweets to monitor the occurrence of COVID-19. Our visualization tool updates weekly and comprehensively analyzes information related to COVID-19 symptoms, case distribution, reinfections, recoveries, and long-term effects based on large language models (LLMs). Figure 1 shows our research objectives: (1) We built a publicly available database of over 9,836,206 COVID-19-related tweets, and this database is set to automatically update with newly collected data weekly. (2) LLMs were built to automatically filter the tweets of self-reporters and extract their mentioned symptoms and recovery cycles. (3) We built a weekly refreshing visualization website, “Covlab,” [30] to track and analyze infection, reinfection, recovery, and long-term effects of COVID-19.

Figure 1. Workflow of “Covlab” online tool.



Data collection: Collected COVID-19-related tweets and manually labeled some tweets as the training set. Models training: Model selection and training were based on annotated datasets to identify the most optimal performing model. Subsequently, self-reported COVID-19 infection tweets were identified from a pool of tweets related to COVID-19. Tracking: Long-term tracking of all individuals who self-reported COVID-19 infection in extracted tweets mentioned symptoms, recovery cycle, long-term effects, and geographical location information from their tweets. Visualization website: Displayed the results of the above analyses.

Methods

Data Collection and Preprocessing

We collected and processed tweet data from January 2020 to April 2024 based on COVID-19-related keywords and hashtags through the Twitter streaming application programming interface (API). The searching keywords included “I.* tested[ed] positive for [covid | coronavirus | covid19 | covid-19],” “My.* [covid | coronavirus | covid19 | covid-19].* symptoms,” “#COVID” “#LongCovid” (All keywords can be found in Multimedia Appendix 1). The following preprocessing operations were conducted on the tweets collected based on keywords. We first converted all words in the tweets to lowercase, standardized the tweets to ASCII encoding using the Unicode library, and tokenized the tweets. Next, we removed all Unicode symbols and punctuation marks, some uninformative characters about usernames such as “@username,” all digits and line breaks in the tweets, all URLs such as “http://,” and all words contained in our stop word library (Multimedia Appendix 2). We converted emoji expressions into their corresponding textual expressions. To provide sufficient datasets for subsequent model training, we constructed a self-reported COVID tweets dataset with manual labeling by 9 native English-speaking annotators after obtaining ethical approval. We also established a set of calibration criteria as shown in Table 1 to ensure the consistency and reliability of the tweet annotations. Two additional annotators conducted a secondary annotation on the labeled data. If their new annotations were inconsistent with the originals, all annotators would decide through a voting process whether the tweet should be classified as a self-reported COVID-19 positive tweet. We also developed a web-based annotation tool (Multimedia Appendix 3) to improve the efficiency of manual annotation. We annotated a total of 105,214 tweets, of which 13,701 were positive samples that described self-reported positive COVID-19 cases and 101,513 were negative samples either not describing self-infection with COVID-19 or irrelevant. Multimedia Appendix 4 presents the types of tweets targeted in our study. The tweet depicted on the left serves as a prototypical instance, delineating self-reported information pertaining to COVID-19 infections. This includes the date of diagnosis and a detailed account of the symptoms experienced by the individual. Conversely, although the tweet shown on the right also references a diagnosis of COVID-19 and details associated symptoms, it diverges from our criteria for target tweets because it recounts a diagnosis pertaining to a third party, specifically a friend of the tweeter, rather than a self-reported account. Therefore, it falls outside the scope of our target dataset.

Table 1. Labeling criteria.

Index	**Annotation Guideline**	**Description**
1	Self-reported Infection	Tweets must be a personal account by the author regarding their experience of contracting COVID-19. If the tweet mentions someone other than the author being infected, such as friends, family, or others, it should be labeled as a negative sample.
2	First-person Narrative	Tweets should use first-person pronouns (eg, “I,” “me,” “my”) to describe the author’s experience of contracting COVID-19, not that of others.
3	Concrete Information	Tweets should provide concrete details, including infection timeline, test results, medical treatments, etc., rather than general discussions or speculations.
4	Symptom Description	Tweets should contain the patient’s personal descriptions of COVID-19 symptoms experienced, such as fever, cough, or difficulty breathing.
5	Confirmation Information	Tweets should mention how the author was confirmed to have contracted COVID-19, such as the type of test conducted (PCR, rapid antigen test), confirmation by a doctor, or official institution validation.
6	Treatment Experience	Tweets should describe the author’s experience of treatment or recovery after self-contracting COVID-19, including isolation, medication, deterioration, or improvement in their health.
7	Infection Timeline	Tweets should contain exact time points or time ranges of the infection rather than general discussions. Providing precise timing helps verify the author’s infection period.
8	Test Results	Tweets should reference the author’s COVID-19 test results, such as testing positive or negative or other relevant test outcomes.
9	Medical Facility	Tweets should mention whether the author received treatment or underwent COVID-19-related tests at a medical facility such as a hospital or clinic.
10	Social Distancing Measures	Tweets should discuss the author’s adoption of social distancing measures due to their infection, such as self-isolation or notifying close contacts.
11	Substantial Evidence	Tweets should contain substantial evidence, such as medical records, official notices, or other documents validating the author’s COVID-19 infection.
12	Exclusion of Transmission	Tweets emphasizing that the author did not transmit the virus to others may indicate a self-reported infection.
13	Consensus Annotation	If three or more annotators provide inconsistent annotations for a particular tweet, a discussion among all annotators should be initiated to reach a final consensus.

Self-Reported COVID-19 Positive Model

To select text-classification models to determine whether a tweet is a self-reported COVID-19-positive tweet, we trained both traditional machine learning methods and fine-tuned LLMs. We divided the manually labeled dataset in which 80% was used as the training set, 10% as the validation set, and the remaining 10% as the test set. We employed word frequency, term frequency-inverse document frequency (TF-IDF) vectors, and feature hashing vectors as methods for text feature extraction, and we adopted 10-fold cross-validation to ensure the reliability of the results.

Because the amount of the manually annotated data was relatively small and the text data features were relatively simple, we experimented with machine learning methods such as Naive Bayes (NB), support vector machines (SVM), and logistic regression (LR) methods. In the SVM method, we utilized the radial basis function (RBF) as the kernel function and set the penalty parameter to 1.2. We employed an L2 regularization as the penalty term in logistic regression, with a regularization strength parameter (C) set to 1.0

As for LLMs, we used BERT [31], RoBERTa [32], XLNet [33], GPT-2 [34], BLOOM [35], and Llama2 [36] for training. These LLMs are pretrained on large datasets using the masked language model (MLM) objective for BERT and RoBERTa, the permuted language model (PLM) objective for XLNet, and the causal language modeling (CLM) objective for GPT-2, BLOOM, and Llama-2. Because GPT-2, BLOOM, and Llama2 are generative models, they may generate nonuniform results, rendering result interpretation difficult. We borrowed the idea from previous work [37], leveraging latent representations from LLMs for supervised label prediction. Hence, we designed different LLM-based classifiers integrated with various LLMs and fully connected neuro networks. Precisely, each LLM serves as a backbone for encoding the tweets instead of generating text. A fully connected neuro network was integrated on top of each LLM for stably accurate detection of self-reported COVID-19 cases. Unlike traditional machine-learning methods requiring feature preprocessing, the LLM-based classifiers take only the text of a tweet as input.

To prevent catastrophic forgetting and ensure that LLMs have a broad understanding of self-reported COVID-19 cases during the training stage, we utilized low-rank adaptation (LoRA) [38] to fine-tune the LLM-based classifiers. LoRA enables the parameters of a model to learn effectively by introducing trainable rank decomposition matrices into the transformer architectures in the LLMs. To achieve this reparameterization, we modified the projection matrices of query, key, value, and feedforward network modules within the self-attention mechanism of the transformer.

The LLM-based classifiers were trained end-to-end with the AdamW [39] optimizer with a cross-entropy objective function. During the training stage, the parameters introduced by LoRA within the pretrained model were updated with gradients, and all remaining parameters were frozen. Early stopping to monitor the accuracy of the validation dataset was implemented during training. All runs were trained on the Nvidia A100 GPU with a batch size of 5 for Llama-2 and 32 for other models.

We evaluated the performance of each model and chose the one that achieved the best results for predicting the daily number of self-reported confirmed cases. Subsequently, we applied a named entity recognition (NER) [40,41] method to extract essential symptom-related keywords from the tweets. For the definition of symptoms, we referred to the descriptions of symptoms within the systematized nomenclature of medicine clinical terms (SNOMED CT)[42] as shown in Table 2. To make our system operational, we deployed the trained model on a server. We employed a script program to continuously collect COVID-19-related tweets from Twitter. The collected tweets were then analyzed using the deployed model, and the results were displayed weekly on the Covlab website. This provided users with up-to-date, analyzed information regarding COVID-19 development.

Table 2. Symptoms and their expression found in self-reported tweets.

Symptoms	Descriptions (with their IDs in CT)
Fever	Febrile (386661006), fever (386661006), feverish (103001002), mill fever (85761009), hyperthermia (1197782006), hay fever (21719001), degrees Fahrenheit (258712004), temperature (722490005), high temperature (285717004), high body temperature (50177009), body temperature above reference range (50177009), increased body temperature (50177009), elevated temperature (50177009), raised temperature (50177009), increased skin temperature (17038008), feeling hot (373932008), feeling hot and cold (103002009), feeling hot and sweaty (373939004)
Chills	Chills (43724002), chills and fever (274640006), chillness (43724002), shivering (43724002), shivering or rigors (248456009), rigor (38880002), brass founders’ ague (74800004), algor (425681008), shakes (26079004), shaking (26079004), trembling (267079009), cold (82272006), head cold (82272006), freeze (48103003), freezing (48103003), frigid (48103003)
Sweating	Sweat (74616000), sweats (415690000), sweating (415691001), cold sweat (83547004), hot

	sweat (224962007), hemopoiesis (445961003), hidrosis (415691001), diaphoresis (52613005), perspiration (415691001), perspire (74616000), perspire profusely (74616000), started to perspire (74616000), perspire all over (74616000), perspire during sleep (74616000), excessive sweating (52613005), profuse sweating (52613005), sweating profusely (52613005)
Runny nose	Sniffle (275280004), nose running (267101005), running nose (267101005), nose dripping (267101005), nasal discharge (267101005), snotty (267101005)
Nasal congestion	Nasal congestion (68235000), congested nose (68235000), stuffed-up nose (68235000), congestion (85804007), stuffed up nose (68235000), stuffed nose (68235000), rhinobyon (68235000), nasal obstruction (232209000), nasal airway obstruction (232209000), stuffiness (232209000)
Nosebleed	Nosebleed (249366005), nose bleeds (249366005), nose bleeding (249366005), bleeding from nose (249366005), nosebleed (249366005), nasal haemorrhage (249366005), epistaxis (249366005), nasal hemorrhage (249366005)
Cough	Cough (49727002), coughing (49727002), nonproductive cough (11833005), hacking cough (59994004), tussiculation, dry cough (11833005), persistent cough (284523002 acute cough (49727002), bad cough (49727002), coughing all night (161933007), evening cough (161933007), morning cough (161932002), coughing up blood (66857006), coughing and deep breathing (371605008), begma (49727002)
Headache	Headache (25064002), migraine (37796009), sick headache (193028008), tension-type headache (398057008), cluster headache syndrome (193031009)
Sneezing	Sneeze (76067001), sneezing (76067001), sneezes (76067001), sneezing symptom (162367006), sternutation (76067001), niesen (76067001), achoo (76067001)
Eye pain	Eye pain (41652007), pain in eye (41652007), ocular pain (41652007), ocular headache (86925001), ocular dryness (162290004), dry eye (162290004), cephalalgia (25064002), diplopia (24982008), double vision (24982008), eyelid edema (89091004)
Loss of taste or smell	Smell (397686008), taste (397627001), lost sense of smell (44169009), absent smell (44169009), sense of smell absent (44169009), anosphrasia, anosmia (44169009), can’t smell (44169009), smell nothing (44169009), disorder of taste (399993004), loss of taste (36955009), absence of sense of taste (36955009), ageusia (36955009)
Sputum	Sputum (45710003), expectoration (45710003), productive cough (28743005), productive cough–green sputum (161924005), productive cough–yellow sputum (161925006), phlegm (52024008)
Shortness of breath	Respiratory disorder (50043002), respiratory disease (50043002), breath (11891009), shortness of breath (267036007), dyspnea (267036007), short breath (267036007), breathless (267036007), difficulty in breathing (230145002), breathing difficult (230145002), labored breathing (248549001), difficulty breath (230145002), breathing painful (75483001)
Sore throat	Sore throat (267102003), throat sore (162388002), throat pain (162388002), pain in throat (162397003), pain throat (162397003)
Dizziness	Dizzy (267102003), throat soreness (267102003), dizzy spells (315018008), dizziness (404640003)
Intolerance to light	Intolerance of light (1285284009), photophobia (409668002), eye sensitive to light (1285284009), light sensitive (1285284009), light sensitivity (1285284009)
Hearing findings	Pains of ears (301354004), ear ache (301354004), otalgia (301354004), earache (301354004), ear aches (301354004), ears pop (162346006), popping sensation in ear (162346006), tinnitus (60862001), noise in ears (60862001)
Loss of appetite	Poor appetite (64379006), decrease in appetite (64379006), inappetence (64379006), lost my appetite (64379006), loss of appetite (79890006), no appetite (79890006), anorexia (79890006), off food (79890006)
Hallucinations	Hallucinations(7011001), hallucination (7011001), illusion(5152006), illusionary (5152006), auditory hallucination (45150006), visual hallucination (64269007), see things (64269007)

Phase Difference Calculation

Because it took time to provide the official report, the COVID-19 occurrence derived from Twitter (predicted curve) was expected ahead of the official report dates (actual curve). Our research used the Hilbert transform (HT) [43] method to calculate the phase difference between the actual and predicted curves. HT is a method used for analyzing time-varying signals [44]. It can transform a real-valued signal into a complex-valued signal, rendering it convenient for phase analysis. In signal analysis the HT is often used to calculate the instantaneous phase of a signal, which can be used to compare the phase difference between two signals. HT is usually more accurate and stable than the traditional Fourier transform method in calculating phase differences. The phase spectrum of the Fourier transform may have discontinuous jumps in some cases, which can lead to incorrect results when calculating phase differences. HT can accurately calculate the instantaneous phase of a signal and avoid this problem.

The daily predicted cases curve $f(t)$ and the daily actual cases curve $g(t)$ share the same time sequence. To calculate the phase difference between them by performing the HT, we can first transform them into their complex-valued signals:

$$H(f(t))=\frac{1}{\pi}P.V.\int_{-\infty}^{\infty}\frac{f(\tau)}{t-1}d\tau \qquad (1)$$

$$H(g(t))=\frac{1}{\pi}P.V.\int_{-\infty}^{\infty}\frac{g(\tau)}{t-1}d\tau \qquad (2)$$

Here, $[H]$ represents the HT operator, and i is the imaginary unit. We can then calculate the instantaneous phase of each signal, usually using the $\arctan$ function to compute the phase angle:

$$Phase(H\{f(t)\})=\arctan\left(\frac{\Im[H\{f(t)\}]}{\Re[H\{f(t)\}]}\right) \qquad (3)$$

$$Phase(H\{g(t)\})=\arctan\left(\frac{\Im[H\{g(t)\}]}{\Re[H\{g(t)\}]}\right) \qquad (4)$$

Finally, we subtract the phase angles of the two signals to obtain the phase difference $\Delta\phi(t)$:

$$\Delta\phi(t)=Phase(H(f(t)))-Phase(H(g(t))) \qquad (5)$$

In addition to calculating the phase difference between the two curves, we also conducted stationarity tests on both curves. Stationarity verification is an important step in time series analysis and is used to determine whether a time series is stationary. We utilized 3 methods, the augmented Dickey-Fuller (ADF) test [45], the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test [46], and the Phillips-Perron (PP) test [47], to verify the stationarity of the two curves. We employed the TimesNet [48] approach from prior research to predict the current trends in COVID-19 development based on the time-series relationships between the self-reported case numbers and the actual case numbers.

Evaluation of the Model

Our evaluation of the model employed the following methodology. True positive (TP) is the number of correct predictions in positive samples, false positive (FP) is the number of incorrect predictions in positive samples, true negative (TN) is the number of correct predictions in negative samples, and false negative (FN) is the number of incorrect predictions in negative samples. Precision is the proportion of positive predictions in all positive samples. Precision is defined as follows:

$$Precision=\frac{TP}{TP+FP} \quad 6$$

Recall is the proportion of correct predictions in the total samples. Recall is defined as follows:

$$Recall=\frac{TP}{TP+FN} \quad 7$$

Accuracy is defined as the percentage of correctly predicted results out of the total sample.

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \quad 8$$

F1-score is defined as follows:

$$F1-score=\frac{2*Precision*Recall}{Precision+Recall} \quad 9$$

To measure the performance under the unbalanced data distribution, in this work we employed the precision-recall (PR) curve and receiver operating characteristics (ROC) curve to display the performance. The ROC curve is a curve of sensitivity versus 1 – specificity on all possible prediction thresholds. Similarly, the PR curve plots precision versus recall on all possible prediction thresholds. In addition, average precision (AP) and area under the curve (AUC) derived from the PR curve and the ROC curve are also generated for quantitative comparisons in this work.

Results

Model Performance for Self-reporting COVID-19 Cases

We evaluated the models with AUC, AUC-PR, accuracy, precision, recall, and F1-score. RoBERTa and BERT achieved the best performance with an AUC of 0.98 and an AP of 0.97 as shown in Figure 2. Notably, all LLMs outperformed traditional machine learning models in AUC and AP, exhibiting an AUC gain from 0.01 to 0.10 and an AP gain from 0.07 to 0.09. According to the benchmark results in Table 3, BLOOM performed the best compared with other models in accuracy and precision with 0.948 and 0.941 whereas the SVM outperformed others in recall and F1-score with 0.9362 and 0.9329, respectively. Combining AUC, accuracy, and recall, we believe that the BLOOM model has the best performance. Subsequently, we utilized the trained model to assist us in selecting self-reported positive tweets from all COVID-19-related tweets.

Table 3. Performance comparison of various machine learning models and LLMs.

	Accuracy	Precision	Recall	F1-score
Machine learning				
NB	86.86%	65.62%	52.43%	58.29%
LR	92.57%	91.73%	92.64%	92.18%
SVM	93.62%	92.96%	*93.62%*	*93.29%*
Large language models				
BERT	93.80%	92.50%	91.00%	91.70%

RoBERTa	93.60%	91.20%	91.80%	91.50%
XLNet	93.00%	90.40%	91.30%	90.80%
GPT-2	94.30%	92.00%	91.00%	92.60%
BLOOM	*94.80%*	*94.10%*	92.00%	93.00%
Llama-2 (7b)	94.20%	93.30%	91.30%	92.30%

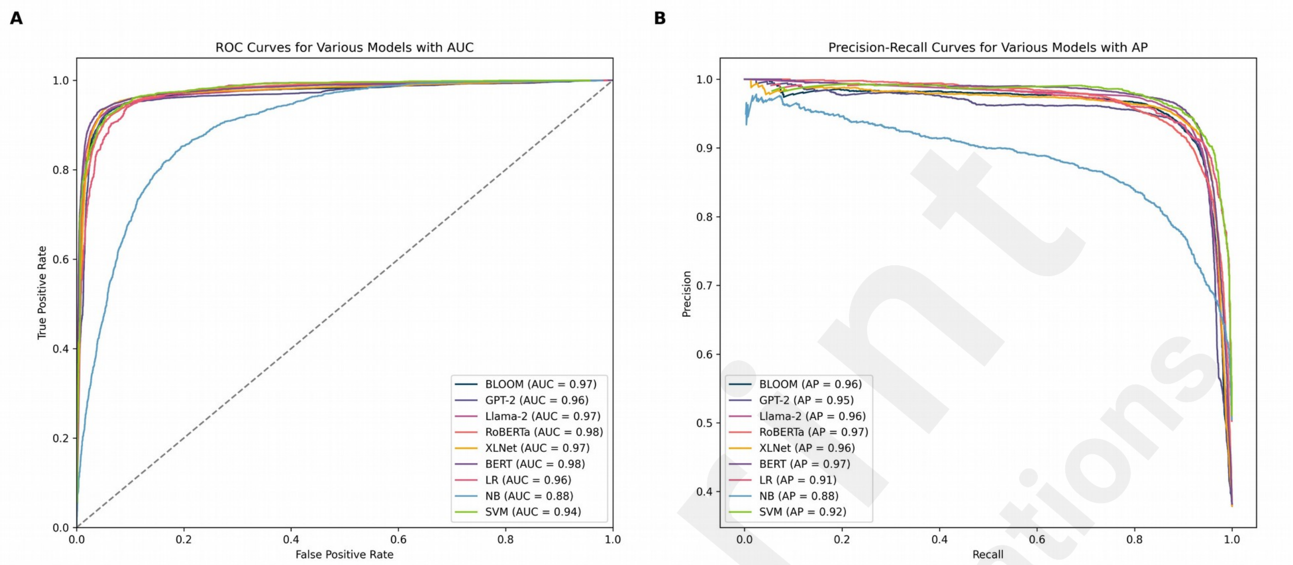


Figure 2. Performance of various machine-learning methods. (A) Receiver operating characteristics (ROC) curve with area values. The area under the curve (AUC) values, provided in the legend, quantify the overall performance of the models, with higher values indicating superior discriminative ability. (B) Precision-recall curve with average precision. LR represents logistic regression; SVM represents support vector machines; NB represents Naive Bayes.

Predicted Trend of COVID-19 Cases

From January 1, 2020, to April 1, 2023, we used the trained BLOOM model to evaluate all the collected COVID-related tweets and filtered out the tweets that the model predicted as self-reported positive cases of COVID-19. Among 7.3 million COVID-related tweets, we identified 317,500 self-reported tweets. Using unique user IDs, we considered multiple tweets reporting a COVID-19 diagnosis by the same user to be a single case, resulting in 262,278 unique self-reported cases. The IDs of these unique confirmed users have been stored separately in an in-house database named the COVID-19 patient database (CPD), and the daily predicted number of confirmed cases by the model has been stored according to coordinated universa time.

To compare the predicted daily case counts with the actual daily case counts, we obtained the actual daily case counts from public platforms such as Johns Hopkins University and *The New York Times*. Then we plotted the actual daily case count and predicted case count on a curve, as shown in Figure 3. The blue line represents the daily actual case counts, and the red line represents our predicted case counts. The red text in the figure describes key events during the outbreak, and the brown text represents the time the variant appeared. We utilized the Hilbert transform method to calculate the phase difference between the 2 curves and found that the predicted curve was leading the actual curve by approximately 7.63 days. The ADF test results indicated values below the critical values of 1%, 5%, and 10%, accompanied by simultaneous *P*-values of 0.0001 and 0.000064, which rejects the hypothesis of the existence of a unit root. Additionally, both the PP and KPSS tests exhibited *P*-values lower than 0.05. Collectively, the outcomes from these 3 tests consistently pointed toward the smoothness of the time series under scrutiny as shown in Multimedia Appendix 5.

There are two distinct peaks in the red curve. We examined the data for the first peak of the predicted curve on October 2, 2020, and the second peak on November 21, 2020. We found that the first peak on October 2, 2020, was due to then-US President Trump's tweet announcing his positive COVID-19 test result, which triggered many Twitter users to also report their self-diagnosed cases on Twitter. On that day, there were 1,495 tweets related to self-reported positive cases. As for the second peak in the prediction curve, we examined the relevant tweets on that day and found that most of them were related to the US election results. Many users tweeted about their infection status and discussed the US epidemic-prevention policies. There were 973 tweets on that day regarding self-reported cases of COVID-19 infection out of all tweets.

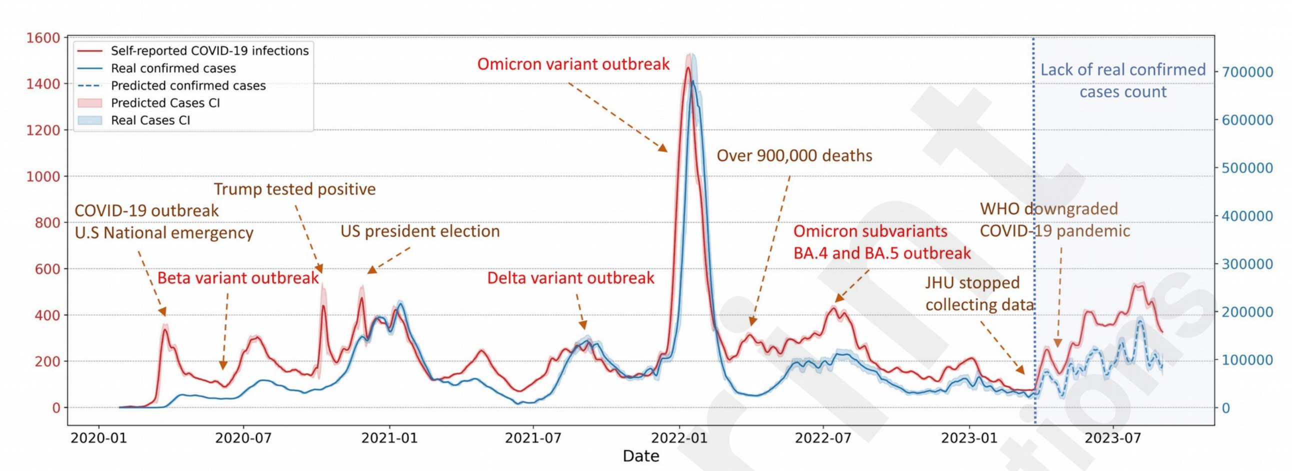


Figure 3. Real cases and predicted cases curves. The blue curve represents the actual daily confirmed cases, and the red curve represents the daily predicted cases. The shaded areas above and below the red and blue curves represent the confidence intervals (CI). The red text represents key events during the outbreak, and the brown text represents the time the variant appeared. The blue shaded area on the right side represents the period during which actual confirmed case data are missing. The solid red line represents the daily self-reported COVID-19 infection numbers, and the blue dotted line represents the predicted actual infection numbers.

Symptoms

We extracted the historical tweets of users from the CPD utilizing their unique user IDs. The temporal scope of these tweets spanned a period from 1 month preceding the self-reported date of symptom onset to 9 months subsequent to the diagnosis date, encompassing a total duration of 10 months. Within a cohort of 24,316 reporting COVID-19 symptoms, an analysis of historical tweets identified the top 10 symptoms. Figure 4(A) plots the temporal frequency of COVID-19 symptom mentions, providing insights into how symptom prevalence evolved over time. Notably, "fever," "headache," "fatigue," and "cough" emerged as consistently common symptoms. The trends observed in these symptoms closely parallel the overall trend in confirmed COVID-19 cases. We also identified that the onset of symptoms like "loss of taste or smell" and "shortness of breath" first became prominent in September 2020, possibly correlating with the emergence of the Beta variant. Similarly, the prevalence of 'sore throat' spiked in late 2021, potentially aligning with the rise of the Omicron variant. The symptom "difficulty breathing" maintained a steady presence across the timeline. Notably, less commonly reported symptoms such as "hallucinations" and "eye pain," not currently recognized by the Centers for Disease Control and Prevention (CDC), appeared sporadically in user reports, suggesting their rarity in COVID-19 cases. As shown in Figure 4(B), these were fever (47.76% mentioned), headache (34.33%), cough (32.84%), shortness of breath (25.37%), generalized body aches (25.37%), difficulty breathing (23.88%), fatigue (23.88%), disorder of smell and/or taste (22.39%), sore throat (20.9%), and eye pain (19.4%), listed in descending order. Notably, all symptoms except for "eye pain" aligned with those recognized by the CDC. Additional symptoms, such as "lethargy" (8.95%), "dizziness" (5.97%), and "hallucinations" (4.47%), although mentioned by a minority, are not currently acknowledged as COVID-19 symptoms by the CDC.

In the dataset of historical tweets from diagnosed individuals, we observed instances in which a patient mentioned multiple symptoms concurrently. To quantify this, we calculated the frequency of co-occurrence of any 2 symptoms and constructed a dependency graph to illustrate these relationships. Figure 4(C) elucidates the correlations among various symptoms, highlighting that most infected individuals reported experiencing a constellation of related symptoms, such as headache, cough, and fever. Furthermore, Figure 4(D) presents a heatmap that visualizes the Pearson correlation coefficients [49] among these symptoms, offering a quantitative view of their interdependencies. To visually represent the range of self-reported symptoms, we utilized a word cloud in Figure 4(E). This graphical representation provides an immediate overview of the symptomatology as expressed by the users in our dataset.

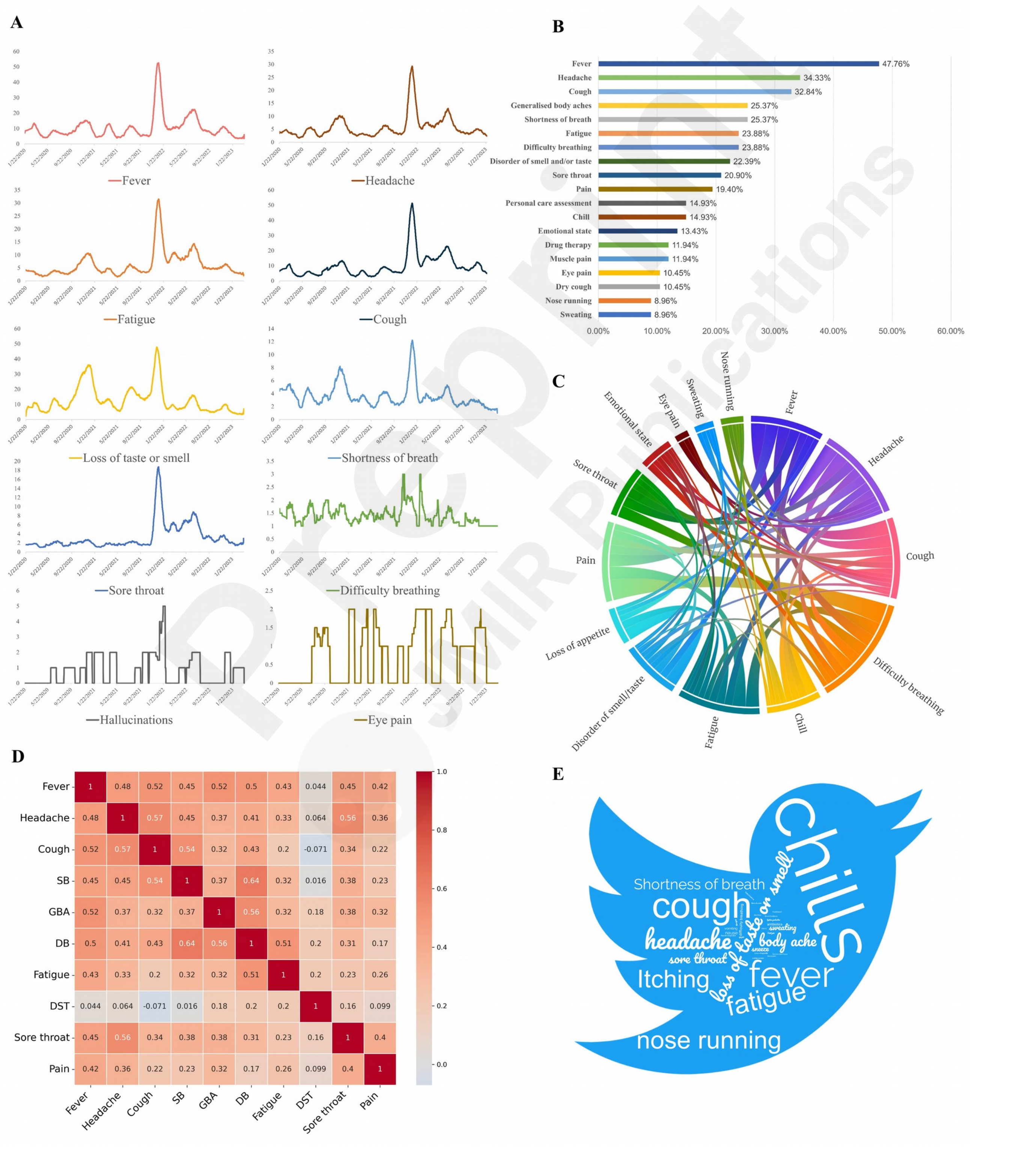


Figure 4. All symptoms mentioned by self-reporting tweets and the correlations among symptoms. (A) represents the number of mentions of COVID-19 symptoms in self-reported tweets over time. (B) represents the percentage of symptoms in all self-reporting COVID tweets, and multiple symptoms can be mentioned in one tweet. Abbreviations: shortness of breath (SB), generalized body ache (GBA), difficulty breathing (DB), disorder of smell/taste (DST). (C) represents the correlations among symptoms mentioned by the same user. The width of the line between two symptoms represents the number of tweets that mention both symptoms. (D) represents a heat map of Pearson correlation coefficients among symptoms. (E) represents the symptoms word cloud. The larger the font size, the greater the number of cases mentioning that symptom.

Reinfection and Rehabilitation

We analyzed historical tweets from users who self-reported COVID-19 infections, identifying 723 individuals who shared their recovery experiences. The annual breakdown of these individuals is as follows: 174 in 2020, 163 in 2021, 135 in 2022, and 251 in 2023. The duration of recovery was primarily inferred from the period mentioned in their tweets. In instances in which the recovery period was not explicitly stated, we computed it by calculating the interval between the date of confirmed diagnosis and the date of reported recovery. The data on self-reported recovery durations have been depicted through the Kaplan-Meier recovery curve [50], as illustrated in Figure 5(A). This graphical representation reveals a gradual decrease in recovery time for COVID-19 patients from 2020 to 2023. Specifically, in 2020, most patients reported a recovery period of around 30 days, with very few recovering in less than 30 days. In contrast, by 2023, the trend had shifted significantly, with most individuals reporting a recovery within approximately 12 days, despite a minority still experiencing recovery periods extending beyond 30 days. Figure 5(B) presents a comprehensive overview of the evolution of recovery periods from 2020 to 2023. Additionally, this figure suggests that the prevalent COVID-19 cases in 2023 were predominantly mild, indicating a possible decrease in the virulence of the virus over time.

In this study we defined a recurrent COVID-19 infection in an individual as a self-reported reinfection occurring more than 30 days after the initial confirmed positive diagnosis. We meticulously tracked the historical tweets of all confirmed patients in CPD. Figure 5(C) presents the distribution of the intervals between the first and second infections alongside the corresponding case counts. This analysis reveals a relatively low likelihood of a repeat infection within 180 days. Most patients who had recovered from an initial infection reported a second infection approximately 260 days later. Moreover, repeat infections occurring between 300 and 600 days postrecovery were also relatively frequent. The longest interval between repeat infections documented in our study extended to 720 days. Of 262,278 patients who self-reported a positive COVID-19 test result, 238,993 indicated a single infection event, and 17,906 reported 2 infections. A smaller subset, comprising 3,283 individuals, reported 3 infections; 1,025 indicated 4 infections; 445 reported 5 infections; 201 individuals reported 6 infections; and 114 indicated 7 or more infections. Remarkably, the highest number of reported reinfections was 9, with 7 individuals documenting their ninth infection. Figure 5(D) shows that among the 238,993 patients with a single infection, 17,906 (7.49%) reported a second infection. A further breakdown shows that 3,283 (1.37%) reported a third infection, 1,025 (0.43%) reported a fourth infection, and 1,071 (0.45%) reported experiencing 5 or more infections. We also performed a statistical analysis of the time intervals between infections among users with multiple infections.

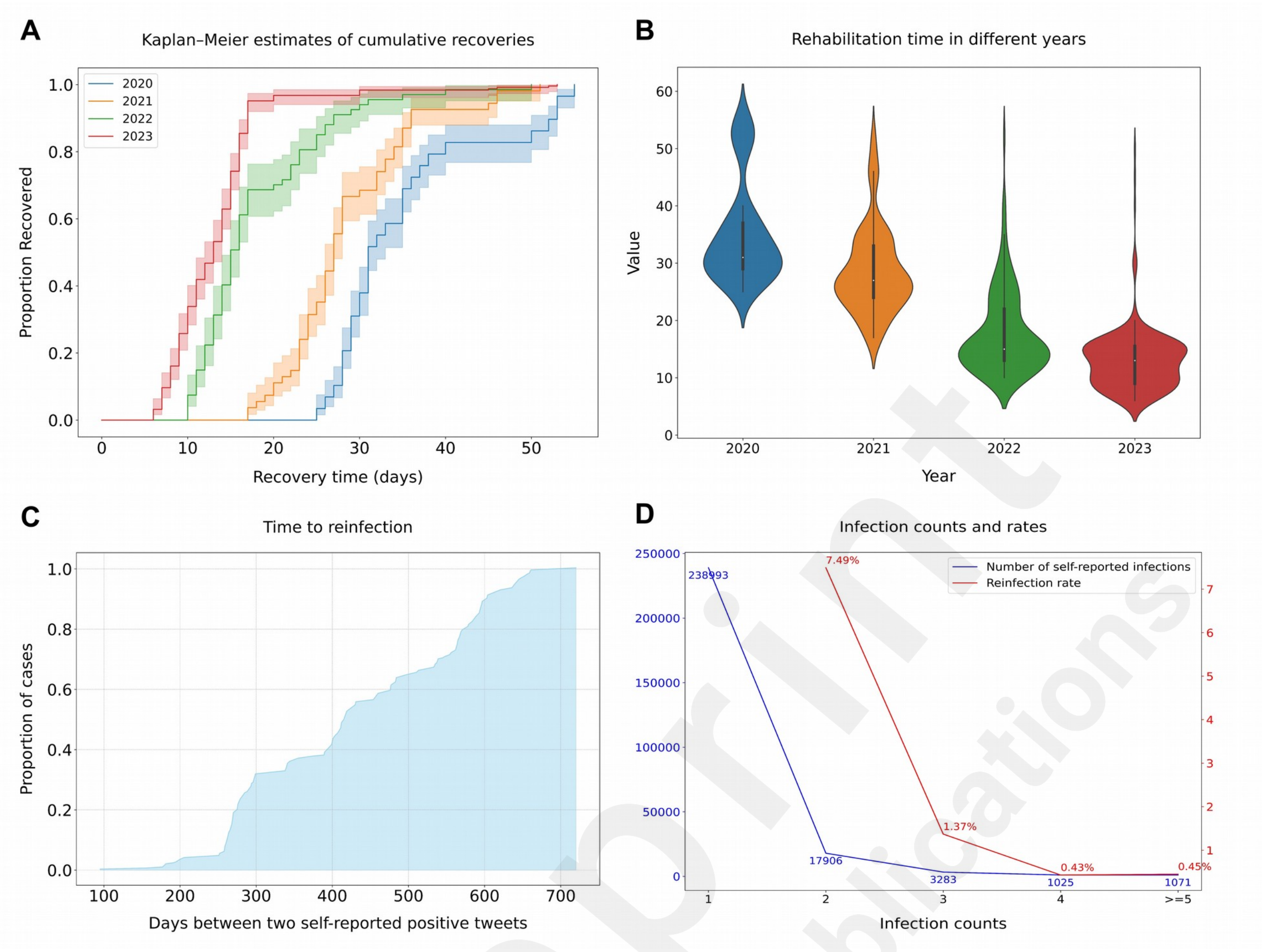


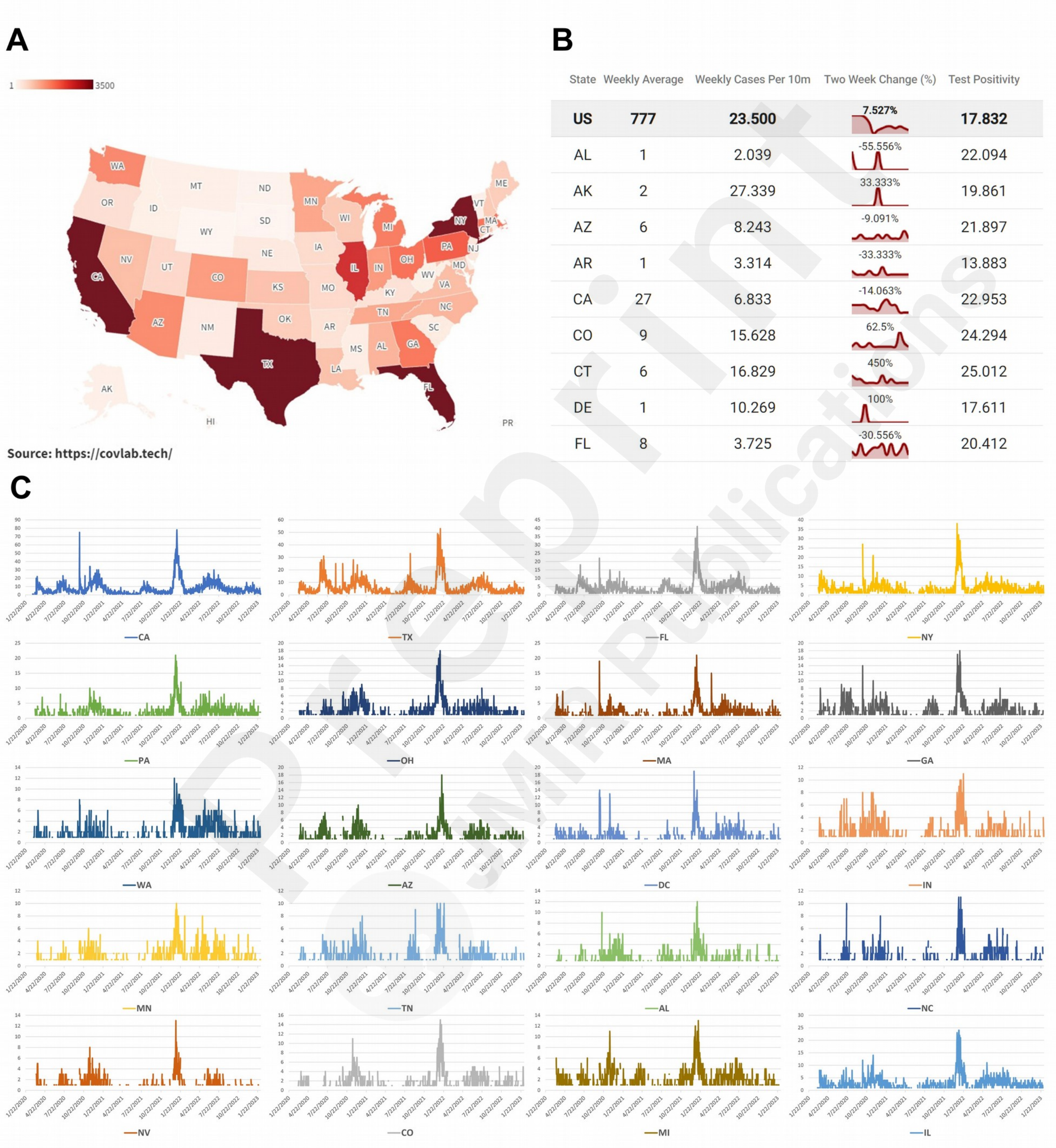
Figure 5. Overview of reinfections and recovery. (A) Kaplan-Meier estimates of cumulative recoveries. (B) Rehabilitation days in different years. (C) Time to reinfection for 238,993 individuals. (D) Reinfection cases and rates.

Distribution of Cases

We extracted the geographic locations of the diagnosed users in CPD, and the distribution of all confirmed patients is shown in Figure 6(A). California had the highest number of self-reported COVID-19 cases, with 8,762 users, followed by Texas (6619 cases), Florida (4245), New York (3566), Illinois (2649), Pennsylvania (2032), Ohio (1868), Massachusetts (1793), Georgia (1785), and Michigan (1677), in descending order. Alabama and Northern Mariana Islands had the least data, with only one self-reported case in each state.

The detailed confirmed cases in each state are shown in Figure 6(B), in which we provide the average case counts for the entire country and each state as well as the number of self-reported cases per 10 million people per week, the trend over the past two weeks, and the positivity rate of each state. For example, we can see that California had the most self-reported COVID-19 cases in the past week, with at least 29 people reporting positive test results. Approximately 7.34 people per 10 million reported self-diagnosed COVID-19-positive status, and the average positivity rate of self-reported cases was 26.69%. However, compared with the previous 2 weeks, the number of self-reported COVID-19 cases decreased by 10.93%. Due to insufficient data, the trend of changes in the past 14 days was unavailable for several states, such as the Virgin Islands and Wyoming.

We also plotted the time-varying curves for the confirmed cases in the top 20 states in terms of confirmed cases, as shown in Figure 6(C). It is evident that the changes in the number of confirmed cases in the top 4 states with the highest number of cases closely resemble the overall trend in the United States. Some states, like Washington, Arizona, Washington, DC, and Indiana, exhibited relatively consistent changes in the number of confirmed cases over time whereas states like Nevada, Colorado, Alabama, and Michigan had less consistent curves, with some dates showing no reported cases. The variation in results could have been influenced by the differing numbers of Twitter users in each state.



Source: https://covlab.tech/

State	Weekly Average	Weekly Cases Per 10m	Two Week Change (%)	Test Positivity
US	777	23.500	7.527%	17.832
AL	1	2.039	-55.556%	22.094
AK	2	27.339	33.333%	19.861
AZ	6	8.243	-9.091%	21.897
AR	1	3.314	-33.333%	13.883
CA	27	6.833	-14.063%	22.953
CO	9	15.628	62.5%	24.294
CT	6	16.829	450%	25.012
DE	1	10.269	100%	17.611
FL	8	3.725	-30.556%	20.412

Figure 6. Cases across the United States and by state. (A) Self-reported number of infections in each state, with darker colors representing higher numbers. (B)Self-reported weekly number of infections, number of infections per 10 million people, change in trend over the previous two weeks, and positivity rate for a subset of states (alphabetical order). (C) Infection curve for the top 20 states with the highest infection numbers.

Covlab Visualization Website

To disseminate timely information to the public and policymakers, we have launched a visualization website, Covlab, which features our trained models and a comprehensive data pipeline. This platform consists of an automated script sequence designed to update the data weekly. The homepage of Covlab offers users real-time access to the total number of tweets collected to date, including those self-reported as COVID-19-positive. On the “Graphs” page, users can explore the most recent weekly growth trends in COVID-19 cases alongside predictive models of actual confirmed cases. This section also enables monitoring of infection trends across various states in the United States as well as the prevailing symptom patterns observed to the current date. Furthermore, Covlab displays the proportions of reinfections and recovery periods, with these metrics also being refreshed weekly. The website’s dynamic graphs and tables serve as a valuable resource, providing the public with up-to-date information on the ongoing evolution of COVID-19, including insights into the symptomatic expressions of emerging variants.

Discussion

Transmission Study

The analysis of self-reported COVID-19 tweets offers a valuable perspective, reflecting the actual progression of the pandemic to a considerable degree. The voluminous data generated by self-reporting individuals on Twitter augment clinical datasets, offering a complementary avenue for long-term observation and tracking. This approach is particularly beneficial in bridging the data gap inherent in in-home self-testing scenarios. Our research indicates that reinfections are relatively common, with the likelihood of multiple infections diminishing over time. The breakdown by state of self-reported COVID-19 tweets facilitates more efficient tracking of disease trends. Furthermore, the approach employed in this study could potentially serve as a general pipeline for researching and analyzing other infectious diseases.

Symptoms and Sequelae Study

Within the subset of Twitter users who self-reported as COVID-19 positive, numerous accounts contained detailed descriptions of symptoms experienced during the infection. These firsthand accounts are a rich source of information. We extracted and analyzed the symptomatology mentioned in these tweets, compiling a list of the most prevalent symptoms reported by patients., Our intriguing findings closely mirror the commonly acknowledged COVID-19 symptoms listed by the CDC. However, some symptoms mentioned by a subset of patients in their tweets are not yet recognized as typical by the CDC. This discrepancy highlights the significant role of our study in identifying potential new symptoms of COVID-19 that have not been widely recognized. Such findings could provide valuable guidance for further clinical investigation into these symptoms and their association with different COVID-19 variants. Additionally, the COVID-19 patient database established through this study offers a robust framework for long-term patient tracking. This database not only complements existing clinical data but also provides an invaluable resource for the study of post-COVID-19 sequelae and long-COVID symptoms, thereby enhancing our understanding of the virus’s long-term impacts.

Limitations and Future Work

Although our method and website are highly useful, there are some limitations. First, Twitter/X has a limited quota of public APIs, which renders the platform hugely expensive to run. Second, potential biases in data collection and the varying distribution of Twitter users across different states may impact the predictive accuracy. Furthermore, external factors, particularly notable events in the United States that occurred in 2020, influenced our prediction outcomes. For instance, the 2 notable peaks in self-reported cases observed in 2020 did not necessarily correlate with an actual increase in infections. Instead, these peaks were primarily driven by external events, prompting a surge in infection reporting on Twitter on those specific days. Another significant constraint is the veracity of the information in tweets. However, despite their questionable reliability, these data offer informative trend analyses and hypotheses valuable for future research and validation. In future work, we will integrate self-reported data from other social platforms such as Reddit to reduce data limits and bias and develop a platform that can predict COVID-19 trends in real time based on self-reported content. We will also apply this pipeline to other infectious diseases that may emerge in the future for understanding and tracking the development and trends of other epidemics.

Data Availability

The aggregated datasets analyzed in this study are available from the corresponding author upon request. Manual labeling data can be viewed through the annotation website (https://labelling.covlab.tech). Guest users can use the username "guest" and password "guest" to log into the system for data access. Actual and predicted daily cases are publicly available on the website (https://covlab.tech).

Author Contributors

JCX and ZYZ collected the data and performed the analyses. JH made the data visualizations. SZ ran the LLM part and wrote the model performance part. GHA, XTT, and XFW accessed the study. LJ and YY provided technical support. JCX drafted the manuscript and designed the website; he and ZYZ wrote the code and conducted all the tests. DX conceived and supervised the study. All authors had full access to all the data in the study as needed and had final responsibility for the decision to submit for publication.

Acknowledgements

This research was supported by National Institute of General Medical Sciences of the National Institutes of Health [R35-GM126985]. We would like to thank all the annotators who participated in the calibration of our data, including Kent Studer, James Tipton, Thadeus Meneses, Noah Berry, Sai Akhil Chopparapu, Trinath Adusumilli, Weinan Zhang, Yijie Ren, Li Su, Chunhui Xu, and Congyu Guo.

Competing interests

None declared.

Multimedia Appendix 1

A list of keywords and hashtags used in data collection.
[DOCX File , 18 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Stop word list.
[DOCX File , 20 KB-Multimedia Appendix 1]

Multimedia Appendix 3

Data preprocessing process and annotation system.
[DOCX File , 3873 KB-Multimedia Appendix 1]

Multimedia Appendix 4

An example of the target cohort.
[DOCX File , 277 KB-Multimedia Appendix 1]

Multimedia Appendix 5

Multiple stationarity verification methods and their results.
[DOCX File , 20 KB-Multimedia Appendix 1]

References

1. https://www.who.int/news/item/05-05-2023-statement-on-the-fifteenth-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-coronavirus-disease-(covid-19)-pandemic.
2. https://coronavirus.jhu.edu/map.html.
3. Banda JM, Tekumalla R, Wang G, Yu J, Liu T, Ding Y, et al. A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. Epidemiologia Multidisciplinary Digital Publishing Institute; 2021 Sep;2(3):315–324. doi: 10.3390/epidemiologia2030024
4. Naseem U, Razzak I, Khushi M, Eklund PW, Kim J. COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. IEEE Transactions on Computational Social Systems 2021 Aug;8(4):1003–1015. doi: 10.1109/TCSS.2021.3051189
5. Müller M, Salathé M, Kummervold PE. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. Frontiers in Artificial Intelligence 2023;6. Available from: https://www.frontiersin.org/articles/10.3389/frai.2023.1023281 [accessed Jun 26, 2023]
6. Alqurashi S, Alhindi A, Alanazi E. Large Arabic Twitter Dataset on COVID-19. arXiv; 2020. doi: 10.48550/arXiv.2004.04315
7. Imran M, Qazi U, Ofli F. TBCOV: Two Billion Multilingual COVID-19 Tweets with Sentiment, Entity, Geo, and Gender Labels. Data Multidisciplinary Digital Publishing Institute; 2022 Jan;7(1):8. doi: 10.3390/data7010008
8. Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi MA, Yang Y-C. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. Journal of the American Medical Informatics Association 2020 Aug 1;27(8):1310–1315. doi: 10.1093/jamia/ocaa116
9. Guo J-W, Radloff CL, Wawrzynski SE, Cloyes KG. Mining twitter to explore the emergence of COVID-19 symptoms. Public Health Nursing 2020;37(6):934–940. doi: 10.1111/phn.12809
10. Wu J, Wang L, Hua Y, Li M, Zhou L, Bates DW, et al. Trend and Co-occurrence Network of COVID-19 Symptoms From Large-Scale Social Media Data: Infoveillance Study. Journal of Medical Internet Research 2023 Mar 14;25(1):e45419. doi: 10.2196/45419
11. Mackey T, Purushothaman V, Li J, Shah N, Nali M, Bardier C, et al. Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter: Retrospective Big Data Infoveillance Study. JMIR Public Health and Surveillance 2020 Jun 8;6(2):e19509. doi: 10.2196/19509
12. Feng Y, Zhou W. Work from home during the COVID-19 pandemic: An observational study based on a large geo-tagged COVID-19 Twitter dataset (UsaGeoCov19). Information Processing & Management 2022 Mar 1;59(2):102820. doi: 10.1016/j.ipm.2021.102820
13. Rusli N, Nordin NZ, Ak Matusin AMR, Yusof JN, Rosley MSF, Ling GHT, et al. Geospatial Mapping of Suicide-Related Tweets and Sentiments among Malaysians during the COVID-19 Pandemic. Big Data and Cognitive Computing Multidisciplinary Digital Publishing Institute; 2023 Jun;7(2):63. doi: 10.3390/bdcc7020063
14. Klein AZ, Magge A, O'Connor K, Amaro JIF, Weissenbacher D, Hernandez GG. Toward Using Twitter for Tracking COVID-19: A Natural Language Processing Pipeline and Exploratory Data Set. Journal of Medical Internet Research 2021 Jan 22;23(1):e25314. doi: 10.2196/25314
15. Sukhavasi N, Misra J, Kaulgud V, Podder S. Geo-sentiment trends analysis of tweets in context of economy and employment during COVID-19. J Comput Soc Sc 2023 Mar 23; doi: 10.1007/s42001-023-00201-2
16. Forati AM, Ghose R. Geospatial analysis of misinformation in COVID-19 related tweets. Applied Geography 2021 Aug 1;133:102473. doi: 10.1016/j.apgeog.2021.102473
17. Chi G, Yin J, Smith ML, Bodovski Y. Global Tweet Mentions of COVID-19. arXiv; 2021. doi: 10.48550/arXiv.2108.06385
18. Guntuku SC, Sherman G, Stokes DC, Agarwal AK, Seltzer E, Merchant RM, et al. Tracking Mental Health and Symptom Mentions on Twitter During COVID-19. J GEN INTERN MED 2020 Sep 1;35(9):2798–2800. doi: 10.1007/s11606-020-05988-8
19. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases Elsevier; 2020 May 1;20(5):533–534. PMID:32087114
20. Wissel BD, Van Camp PJ, Kouril M, Weis C, Glauser TA, White PS, et al. An interactive online dashboard for tracking COVID-19 in U.S. counties, cities, and states in real time. Journal of the American Medical Informatics Association 2020 Jul 1;27(7):1121–1125. doi: 10.1093/jamia/ocaa071
21. Zohner YE, Morris JS. COVID-TRACK: world and USA SARS-COV-2 testing and COVID-19 tracking. BioData Mining 2021 Jan 20;14(1):4. doi: 10.1186/s13040-021-00233-2
22. Reinfection rate in a cohort of healthcare workers over 2 years of the COVID-19 pandemic | Scientific Reports. Available from: https://www.nature.com/articles/s41598-022-25908-6 [accessed Aug 25, 2023]

23. Flacco ME, Acuti Martellucci C, Soldato G, Carota R, Fazii P, Caponetti A, et al. Rate of reinfections after SARS-CoV-2 primary infection in the population of an Italian province: a cohort study. Journal of Public Health 2022 Dec 1;44(4):e475–e478. doi: 10.1093/pubmed/fdab346
24. Rivelli A, Fitzpatrick V, Blair C, Copeland K, Richards J. Incidence of COVID-19 reinfection among Midwestern healthcare employees. PLOS ONE Public Library of Science; 2022 Jan 4;17(1):e0262164. doi: 10.1371/journal.pone.0262164
25. Deng L, Li P, Zhang X, Jiang Q, Turner D, Zhou C, et al. Risk of SARS-CoV-2 reinfection: a systematic review and meta-analysis. Sci Rep Nature Publishing Group; 2022 Dec 1;12(1):20763. doi: 10.1038/s41598-022-24220-7
26. Ren X, Zhou J, Guo J, Hao C, Zheng M, Zhang R, et al. Reinfection in patients with COVID-19: a systematic review. Glob Health Res Policy 2022 Apr 29;7(1):12. PMID:35488305
27. Wu X, Wang Z, He Z, Li Y, Wu Y, Wang H, et al. A follow-up study shows that recovered patients with re-positive PCR test in Wuhan may not be infectious. BMC Medicine 2021 Mar 15;19(1):77. doi: 10.1186/s12916-021-01954-1
28. O Murchu E, Byrne P, Carty PG, De Gascun C, Keogan M, O'Neill M, et al. Quantifying the risk of SARS-CoV-2 reinfection over time. Rev Med Virol 2022 Jan;32(1):e2260. PMID:34043841
29. Bourmistrova NW, Solomon T, Braude P, Strawbridge R, Carter B. Long-term effects of COVID-19 on mental health: A systematic review. Journal of Affective Disorders 2022 Feb 15;299:118–125. doi: 10.1016/j.jad.2021.11.031
30. Covlab:Uncover the connection, explore COVID-19 cases through tweets. Available from: https://covlab.tech
31. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv; 2019. doi: 10.48550/arXiv.1810.04805
32. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv; 2019. doi: 10.48550/arXiv.1907.11692
33. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. Advances in Neural Information Processing Systems Curran Associates, Inc.; 2019. Available from: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html [accessed Jan 10, 2024]
34. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog 2019;1(8):9.
35. Workshop B, Scao TL, Fan A, Akiki C, Pavlick E, Ilić S, et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. arXiv; 2023. Available from: http://arxiv.org/abs/2211.05100 [accessed Jan 10, 2024]
36. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv; 2023. doi: 10.48550/arXiv.2307.09288
37. Li Z, Li X, Liu Y, Xie H, Li J, Wang F, et al. Label Supervised LLaMA Finetuning. arXiv; 2023. doi: 10.48550/arXiv.2310.01208
38. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. arXiv; 2021. doi: 10.48550/arXiv.2106.09685
39. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. arXiv; 2019. doi: 10.48550/arXiv.1711.05101
40. Li J, Sun A, Han J, Li C. A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering 2022 Jan;34(1):50–70. doi: 10.1109/TKDE.2020.2981314
41. Bhatia P, Celikkaya B, Khalilia M, Senthivel S. Comprehend Medical: A Named Entity Recognition and Relationship Extraction Web Service. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) 2019. p. 1844–1851. doi: 10.1109/ICMLA.2019.00297
42. Chang E, Mostafa J. The use of SNOMED CT, 2013-2020: a literature review. Journal of the American Medical Informatics Association 2021 Sep 1;28(9):2017–2026. doi: 10.1093/jamia/ocab084
43. Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc Lond A 1998 Mar 8;454(1971):903–995. doi: 10.1098/rspa.1998.0193
44. Benitez D, Gaydecki PA, Zaidi A, Fitzpatrick AP. The use of the Hilbert transform in ECG signal analysis. Computers in Biology and Medicine 2001 Sep 1;31(5):399–406. doi: 10.1016/S0010-4825(01)00009-9
45. Mushtaq R. Augmented Dickey Fuller Test. Rochester, NY; 2011. doi: 10.2139/ssrn.1911068
46. Shin Y, Schmidt P. The KPSS stationarity test as a unit root test. Economics Letters 1992 Apr 1;38(4):387–392. doi: 10.1016/0165-1765(92)90023-R
47. Breitung J, Franses PH. ON PHILLIPS–PERRON-TYPE TESTS FOR SEASONAL UNIT ROOTS. Econometric Theory Cambridge University Press; 1998 Apr;14(2):200–221. doi: 10.1017/S0266466698142032
48. Wu H, Hu T, Liu Y, Zhou H, Wang J, Long M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. arXiv; 2023. doi: 10.48550/arXiv.2210.02186

49. Benesty J, Chen J, Huang Y, Cohen I. Pearson Correlation Coefficient. In: Cohen I, Huang Y, Chen J, Benesty J, editors. Noise Reduction in Speech Processing Berlin, Heidelberg: Springer; 2009. p. 1–4. doi: 10.1007/978-3-642-00296-0_5
50. Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). BMJ British Medical Journal Publishing Group; 1998 Dec 5;317(7172):1572–1580. PMID:9836663

Abbreviations

ADF test: Augmented Dickey-Fuller test
AP: average precision
API: application programming interface
AUC: area under the curve
CDC: Centers for Disease Control and Prevention
COVID-19: Coronavirus disease 2019
CPD: COVID-19 patient database
FN: false negative
FP: false positive
HT: Hilbert Transform
KPSS test: Kwiatkowski-Phillips-Schmidt-Shin test
LLM: large language model
LoRA: Low-Rank Adaptation
PP test: Phillips-Perron test
PR: precision-recall
ROC: receiver operating characteristics
SNOMED CT: systematized nomenclature of medicine clinical terms
TN: true negative
TP: true positive

Supplementary Files

Multimedia Appendixes

Table 1. A list of keywords and hashtags used in data collection.
URL: http://asset.jmir.pub/assets/b165150bb0cca7ee5dd02385b329e1b6.docx

Table 2. Stop word list.
URL: http://asset.jmir.pub/assets/6d218f3770b6979bc3fe631495d634ae.docx

Data preprocessing process and annotation system.
URL: http://asset.jmir.pub/assets/afca292c25f76898cef9ef9a83019ad9.docx

An example of the target cohort.
URL: http://asset.jmir.pub/assets/ceb4ee6f61a13382f22bab0370610530.docx

Multiple stationarity verification methods and their results.
URL: http://asset.jmir.pub/assets/c41d9479e7e5b7af496f8ed5056bfaf1.docx