Characterizing Public Sentiments and Drug Interactions during COVID-19: A Pretrained Language Model and Network Analysis of Social Media Discourse

Wanxin Li¹, Yining Hua², Yining Hua², Yining Hua², Peilin Zhou⁵, Li Zhou³,⁴, Xin Xu¹*, Jie Yang¹*

¹ School of Public Health, the Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China;
² Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA;
³ Harvard Medical School, Boston, Massachusetts, USA;
⁴ Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital, Boston, Massachusetts, USA;
⁵ Hong Kong University of Science and Technology, Guangzhou, Guangdong, China.

Keywords: COVID-19, Natural Language Processing, Drugs, Social Media, Pharmacovigilance, Public Health

Abstract

Objective: Harnessing drug-related data posted on social media in real time can offer insights into how the pandemic impacts drug use and monitor misinformation. This study developed a natural language processing (NLP) pipeline tailored for the analysis of social media discourse on COVID-19 related drugs.

Methods: This study constructed a full pipeline for COVID-19 related drug tweet analysis, utilizing pre-trained language model-based NLP techniques as the backbone. This pipeline is architecturally composed of

* Co-corresponding authors: Jie Yang (jieynlp@gmail.com), Xin Xu (xuxinsummer@zju.edu.cn)
four core modules: named entity recognition (NER) and normalization to identify medical entities from relevant tweets and standardize them to uniform medication names, target sentiment analysis (TSA) to reveal sentiment polarities associated with the entities, topic modeling to understand underlying themes discussed by the population, and drug network analysis to potential adverse drug reactions (ADR) and drug-drug interactions (DDI). The pipeline was deployed to analyze tweets related to COVID-19 and drug therapies between February 1, 2020, and April 30, 2022.

**Results:** From a dataset comprising 2,124,757 relevant tweets sourced from 1,800,372 unique users, our NER model identified the top five most-discussed drugs: Ivermectin, Hydroxychloroquine, Remdesivir, Zinc, and Vitamin D. Sentiment and topic analysis revealed that public perception was predominantly shaped by celebrity endorsements, media hotspots, and governmental directives rather than empirical evidence of drug efficacy. Co-occurrence matrices and complex network analysis further identified emerging patterns of DDI and ADR that could be critical for public health surveillance like better safeguarding public safety in medicines use.

**Conclusion:** This study evidences that an NLP-based pipeline can be a robust tool for large-scale public health monitoring and can offer valuable supplementary data for traditional epidemiological studies concerning DDI and ADR. The framework presented here aspires to serve as a cornerstone for future social media-based public health analytics.

### 1. Introduction

The emergence of the COVID-19 pandemic has induced an immediate need for effective pharmacotherapies. While the development and application of such therapies are critically important, they are also influenced by an array of political, economic, and social factors. For example, public pronouncements by high-profile
figures, such as former U.S. President Donald Trump's endorsement of hydroxychloroquine, have led to its irrational use and consequential public health crises.[1] Traditional pharmacovigilance mechanisms, reliant on clinical trials and formal reporting systems like MedWatch and DrugBank,[2-4] offer valuable but lagged information. These traditional approaches are plagued by inefficiencies, reporting biases, and a lack of timeliness, thereby lacking comprehensive coverage of the population's sentiments and experiences.[5-8]

In this context, real-time public comments on pharmacotherapies such as medications on social media provide a valuable resource for complementing research on drug use or repositioning for COVID-19. In addition to the fast accessibility, timeliness, and comprehensive population coverage, social media can also supply real-world evidence on how people respond to different drugs, thus helping researchers mine novel drug potency or side effects.[9-11] Social media also offer data on drugs not typically included in pharmacovigilance datasets, such as over-the-counter drugs,[12] herbal remedies,[13] and other non-traditional treatments.[14] However, the sheer volume and noise in social media data require robust computational methodologies for effective analysis.

Natural language processing (NLP) technologies offer a solution to these challenges. Earlier studies, such as the study conducted by Aramaki et al. in 2011, demonstrated that Twitter data could be mined to monitor influenza outbreaks using machine learning and rudimentary NLP techniques.[15] Contemporary research in this domain has benefitted immensely from advancements in deep learning and specialized NLP tools, such more specialized NLP tools for analyzing social media data[16] and public health datasets for social media posts.[17] These have made it increasingly feasible to sift through large volumes of colloquial, noisy text to extract meaningful insights on public health.

Within the realm of COVID-19 drug research, substantial efforts have been made to analyze pharmacotherapy-related topics during COVID-19. For example, Hua et al. utilized BERT models to examine
public perceptions of specific, albeit controversial, medications and found these perceptions to be heavily skewed by misinformation and partisan biases.[18] However, the study suffered from methodological limitations, including a narrow and subjectively chosen selection of drugs (i.e., hydroxychloroquine and ivermectin versus Molnupiravir and Remdesivir). Similarly, Wu et al. made the very first attempts to construct co-occurrence networks to study symptomatology, but their technique was solely based on lexicon matching, thus subject to false positives.[19] There is still a lack of data-driven pipeline with state-of-the-art NLP tools and other big data analysis techniques to automatic extract drug information through social media data from the view of public health.

To address these gaps, this study employs NLP methodologies and network analysis for an extensive assessment of COVID-19 drug-related discourse on social media. We contribute to the existing literature in several ways:

1. Employing deep learning methodologies for named entity recognition (NER), thereby reducing the false positives associated with traditional keyword matching.
2. Re-examining public sentiments and concerns regarding COVID-19 medications, utilizing target sentiment analysis (TSA) and topic modeling.
3. Conducting a comprehensive assessment of adverse drug reactions (ADR) and drug-drug interactions (DDI) through network analysis techniques.

We demonstrate that our integrated NLP pipeline can serve as a robust framework for extracting and analyzing drug-related information, thereby enhancing the scope and effectiveness of social media-based pharmacotherapy analysis.

2. Methods
As shown in Fig. 1, the study workflow is organized in three primary stages: data collection, development of an NLP pipeline, and subsequent data analysis using the constructed pipeline. Initially, we curated a dataset of English tweets related to COVID-19. After a preprocessing phase that excluded tweets with URLs, a NLP pipeline was developed to extract and normalize the drugs/diseases mentioned in these tweets. Finally, we examined the time trends of drug mentions, public sentiment, and discussion topics towards drugs, as well as the co-occurrence network of drug-drug and drug-disease pairs. Ethical approval for this study was granted by the Institutional Review Board of Zhejiang University.

Fig.1 Workflow of drug analysis with NLP on Twitter. NLP: natural language process; CT-BERT: COVID-Twitter-BERT; METS-CoV: Medical Entities and Targeted Sentiments on COVID-19-related tweets, NER dataset containing medical entities and targeted sentiments from COVID-19 related tweets.[17]

2.1 Data collection and preprocessing

The used tweets were obtained from a public dataset provided by Chen et al[20] and were downloaded using
Twitter's API. The downloaded data included full tweet texts and corresponding metadata such as timestamps and user information. Tweets containing URLs were excluded from the analysis, as they often only contained summaries or quotations of the original tweet. The data collection process adhered to Twitter's privacy and data use management policies.

2.2 NLP pipeline development

The NLP pipeline consists of four principal modules: Named Entity Recognition (NER), Target Sentiment Analysis (TSA), topic modeling, and drug network analysis. For the NER and TSA modules, we leveraged state-of-the-art models developed in our prior work Medical Entity and Targeted Sentiment on COVID-19 Related Tweets (METS-CoV). Details on model construction can be found in Fig. A.1.

2.2.1 Named entity recognition (NER) and normalization

The NER model aims to extract drug entities from tweets. The model we developed, CT-BERT-NER, was constructed using the COVID-Twitter-BERT (CT-BERT) framework, a widely adopted language model pre-trained on 160 million COVID-19-related tweets. It was trained on the entire training set of the NER subset of METS-CoV. Upon evaluation, it showed F1 scores of 86.35% for drug entity recognition and 81.85% for symptom entity recognition on the corresponding test set. We used the model trained on all entity types instead of on drug entities only to enable the nuanced differentiation of drug entities from other types of entities.

To standardize colloquial expressions of drugs among the extracted entities, we manually searched from Wikipedia for NER-identified drug entities with frequency more than 1000 to map colloquial drug expressions and their standardized concepts (i.e., drug trade names, chemical names, and generic names). We conducted an accuracy assessment using a random sample of 50 tweets for each of the top five most frequently mentioned drugs and symptoms, as identified through two methods: NER combined with lexicon-based extraction (NER
+ lexicon) and lexicon-based extraction alone. Our results demonstrated that the NER + lexicon method achieved an accuracy rate of 100%, significantly surpassing the 89% accuracy achieved by the lexicon-only approach. Further details on this comparison are available in Table A.1.

2.2.2 Targeted sentiment analysis (TSA)

The TSA module aims to quantify user’s sentiment toward specific drug entities within tweets. The TSA model, CT-BERT-TSA[17], is a three-class model developed based on the BERT-SPC framework.[21] Similar to BERT-SPC, CT-BERT-TSA model treats targeted sentiment analysis as a sentence pair classification task, by appending the identified drug entity to the end of the tweet context and then feeding this sentence pair into CT-BERT model for three-class prediction (i.e., positive, neutral, and negative). Upon testing on the TSA test set of METS-CoV, the model showed an F1 score of 62.67% and an accuracy rate of 75.07%,[17] across four entity types — Person, Drug, Disease, and Vaccine.

2.2.3 Topic model analysis

To discern prevailing public interests in the most discussed drugs, we implemented Latent Dirichlet Allocation (LDA) for topic modeling, utilizing the LdaModel function from the Gensim package.[22] Topic numbers were determined based on conventional evaluation metrics, including low perplexity and high coherence scores.[23] [24] Detailed methodologies are delineated in Fig. A.2.

2.2.4 Drug network analysis

For elucidating potential relationships among drugs, we constructed this drug network analysis module to generate incidence matrices and visualize co-occurrence networks using Gephi.[25] For enhanced comprehensiveness, we incorporated a variant supported by the Anatomical Therapeutic Chemical Classification System (ATC),[26] in addition to the Gephi-based visualization. In addition, we used the NER model to extract symptom entities and normalize them through a pre-summarized lexicon list[27] to extend
our analysis to drug-symptom networks. The constructed networks feature nodes representing either frequently occurring drugs or symptoms. As our focus is not on causal relationships but rather on the interplay between entities, we employed undirected graphs and used semantic cosine similarity[28] as the distance metric.

2.3 Pipeline deployment

Upon completion of the NLP pipeline, we proceeded to its deployment on the pre-processed dataset of COVID-19-related tweets. We first applied the NER and normalization module on the pre-processed dataset to extract and standardize drug entities to drug concepts. Following this standardization, we conducted a distributional analysis of drug mentions to discern time trends, thereby capturing the evolving popularity of these drugs. We also gather related news and the trend of weekly new COVID-19 cases to show a more holistic view of the shift of drug popularity over time. For clarity and simplicity, we only illustrate the top five most discussed drugs.

Subsequently, we used the TSA model to assign each drug entity of the five types a sentiment type. To gain a deeper understanding, we also conducted a time-trend analysis on the positive and negative tweets for the five drugs and visualized the results. Building upon our understanding of public sentiment, we turned to topic modeling via LDA to explore the thematic concentrations in the discourse surrounding these drugs. The model yielded the 20 most probable keywords and bigrams for each identified topic, enabling us to summarize the primary themes. We further analyzed the topic distribution associated with each of the top five drugs.

Finally, we constructed co-occurrence networks for drug-drug and drug-symptom interactions to provide a relational overview that complements our earlier analyses. All drugs with more than 1,000 mentions over time were included in the analysis. Meanwhile we also zoomed in to analyze the five most-discussed drugs.
3. Results

3.1 Data summary and trends of drug mention tweets

This study used a dataset consisting of 471,371,477 COVID-19-related tweets in English, which were collected between February 1, 2020 and April 30, 2022. After excluding tweets containing URLs, the final dataset used for this study consisted of 169,659,956 tweets from 103,682,686 users. Using CT-BERT-NER, we identified 2,124,757 drug-related tweets from 1,800,372 unique Twitter users, accounting for approximately 1.25% of the raw COVID-19-related tweets dataset. Table A.2 (in supplementary materials) provides more detailed statistical results of the medical entity recognition.

Table A.3 presents the 67 most frequently mentioned drugs, each with an occurrence exceeding 1000 times. The most frequent taxonomies are Anatomical Therapeutic Chemical Classification System (ATC)[26] is N (nervous system drugs) and J (anti-infective drug). We ranked the total occurrence of all drugs and identified the top five most-mentioned drugs: Ivermectin, Hydroxychloroquine, Remdesivir, Zinc, and Vitamin D to visualize their weekly time trends. Fig. 2 presents these temporal trends.

Among the five drugs, the public focused mostly on repurposed drugs (i.e., hydroxychloroquine and ivermectin), followed by daily supplements (i.e., zinc and vitamin D). The only officially approved drug among the five, Remdesivir, received the least attention. The frequency of discussion of hydroxychloroquine and ivermectin fluctuated significantly across time, which seemed to be related to relevant news events or policies (marked in Fig.2). In the early stage of the pandemic, drug-related discussions focused on hydroxychloroquine, with two prominent peaks occurring at May 24\textsuperscript{th} and August 2\textsuperscript{nd} of 2020. Discussion of ivermectin began to increase in the later stages of the pandemic, with only one prominent peak located at September 5\textsuperscript{th} of 2021. In contrast, Remdesivir received the least public attention, which increased only
sporadically throughout the pandemic, with no apparent pattern and a much lower peak on May 3rd of 2020. As supplements to COVID-19 treatments, vitamin D and zinc elicited much less public interest than ivermectin and hydroxychloroquine, with no significant outbreaks or visible patterns.

**Fig. 2** Weekly popularity trends of the top five most-mentioned drugs on Twitter examined with COVID-19-related tweets collected between February 1, 2020 and April 30, 2022. The left Y-axis represents the total number of tweets for each drug in a given week (unit: thousand tweets). The right Y-axis represents weekly new case count (unit: million tweets). The new case counts were collected from World Health Organization (WHO)[29] on a weekly basis, beginning on February 1st, 2020. Given that the dataset is confined to English-language tweets, the scope of new case counts was likewise restricted to the top four English-speaking nations with the highest Twitter activity: the United States, the United Kingdom, the Philippines, and Canada.[30]

### 3.2 Changes in sentiment for five most frequent mentioned drugs

We calculated sentiment proportion for the five drugs and the weekly time trends of positive and negative tweets. Fig. 3a shows the visualization of the overall attitude proportions. The public tended to hold positive and neutral attitudes toward the repurposed drugs, ivermectin and hydroxychloroquine. The immune
supplements, zinc and vitamin D, were frequently mentioned with positive sentiments. The only COVID-19 drug approved by the FDA, Remdesivir, received the lowest positive attitude (12.8%), far lower than those of the other drugs.

Fig. 3b and 3c presents weekly trends of tweets expressing positive and negative attitudes, respectively. The major turning points of the trends tend to coincide with new government policies, major social events, and research findings. The criticism of remdesivir (Fig. 3c) and ivermectin increased over time since September 2021, and the turning point for remdesivir came at almost the same time of emerging studies showing that the drug is ineffective[31] and has severe side effects.[32-34] For ivermectin, public sentiment was associated with announcements of health authorities and celebrity effects. For example, the FDA denouncing the use of ivermectin for COVID-19 on August 29th, 2021 had simultaneously increasing negative discussions.
Fig. 3 Sentiment analyses of the five top-discussed drug from February 1, 2020 to April 30, 2022, grouped according to their polarity, including (a) sentiment distribution, (b) weekly ratio of positive tweets, and (c) weekly ratio of negative tweets. The denominator of the percentage was the entities with sentiment.

3.3 Topics distributions of drug mentioned tweets

We applied the LDA topic model to all drug-related tweets and obtained 15 general topics based on their relatively high topic coherence scores and low confusion levels (further discussed in Fig. A.3). We display the corresponding top 20 most likely keywords in Table 1, and assigned a theme for each topic from these keywords. The topic "clinical treatment effect of drugs" dominated the discussions, accounting for 13.60% of
all related tweets. Other popular topics included "physical symptoms" (11.84%) and "causes of death" (9.28%),
closely followed by topics such as "immune response" (8.14%), "general treatment" (8.23%), and "daily
supplement intake" (7.27%).

In addition to the overall topic summary, we explored the distribution of the 15 topics for the five drugs.
Fig. 4 shows a visualization of the distribution. For ivermectin, the prominent theme was "immune response".
In contrast, discussions of remdesivir centered on "hospital care". Hydroxychloroquine received relatively
even attention among the three topics "causes of death", "drug scare", and “COVID control”. Vitamin D was
frequently mentioned in tweets about "daily life", and the main topics about zinc focused on "hospital care"
and "COVID control".

Table 1 Topic model on drug related Tweets. Username and other sensitive information were masked off
using @USER. Public figures such as @realdonaldtrump are shown in their usernames.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Keywords</th>
<th>Example</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical treatment effect of drugs</td>
<td>treatment, covid, study, drug, effective, trial, hydroxychloroquine, prove, safe, covid, evidence, prevent, clinical, vaccine, recommend, cheap, infection, continue, antiviral, efficacy</td>
<td>@USER Ivermectin is pretty safe but the evidence for it being efficacious against SARS-CoV-2 is lacking.</td>
<td>13.60%</td>
</tr>
<tr>
<td>Physical symptoms</td>
<td>day, test, symptom, week, covid, positive, feel, month, steroid, bad, start, med, ago, sick, time, recover, antibiotic, hour, fine, insulin</td>
<td>When I had covid, it was mild fever for a day, &amp; it was gone</td>
<td>11.84%</td>
</tr>
<tr>
<td>COVID control</td>
<td>cure, vaccine, spread, control, lie, people, drug, covid, push, approve, hydroxychloroquine, claim, force, medium, ban, talk, science, government, experimental, president</td>
<td>@USER I agree with you re lockdown. Just not on HCQ and vaccines.</td>
<td>10.36%</td>
</tr>
<tr>
<td>Causes of death</td>
<td>death, people, die, covid, kill, reason, heart, drug, cancer, cocaine, dead, trust, epidemic, rate, bad, create, attack, sound, result, fentanyl</td>
<td>@USER Of course, if they died 30 minutes after taking fentanyl but had a positive covid test, guess what their official cause of death is listed as?</td>
<td>9.28%</td>
</tr>
<tr>
<td>General treatment</td>
<td>doctor, treat, patient, steroid, risk, covid, covid, severe, infection, blood, pill, medication, antibiotic, receive, prescribe, illness, lung, hospitalize, prescription, aspirin</td>
<td>@USER migraines have a very specific causality (i had them for like 20 years), I wonder if the Covid version is one? -- I would try warm compress, NSAIDs, and maybe nasal irrigation with like a neti pot</td>
<td>8.23%</td>
</tr>
<tr>
<td>Immune response</td>
<td>virus, system, body, immune, corona, zinc, fight, antibody, care, immunity, bleach, deficiency, cell, response, kill, covid, infect, injectingdisinfectant, human, boost</td>
<td>In addition, mAbs have been shown to improve survival in patients hospitalized with COVID-19 who have not mounted their own immune response.</td>
<td>8.14%</td>
</tr>
<tr>
<td>Daily supplement intake</td>
<td>vitamin, people, level, covid, eat, healthy, cold, flu, catch, protect, food, hand, low, stay, vit, survive, bad, chance, common, worry</td>
<td>@USER regularly take vitamins and a Vitamin D supplement. I started taking the Vitamin D supplement because I wasn't going outside as much at the start of the pandemic. As soon as started taking the Vitamin D</td>
<td>7.27%</td>
</tr>
<tr>
<td>Public panic</td>
<td>live, save, pandemic, life, start, lockdown, real, people, buy, time, basically start, money, watch, steroid, hit, hard, normal, deadly, cough, break</td>
<td>@USER waste of time, politicians have organised orgy’s with cocaine, male &amp; female hookers, during lockdown! like they give a fuck about a petition hahaha</td>
<td>7.04%</td>
</tr>
<tr>
<td>Hospital care</td>
<td>patient, care, hospital, remdesivir, treatment, cocktail, covid, covid, injection, require, oxygen, admit, medical, lead, ventilator, pay, provide, health, remove, source</td>
<td>Both hospitalized and treated immediately with Oxygen &amp; Remdesivir for covid @ the same time. Both went into heavy psychosis.</td>
<td>5.56%</td>
</tr>
<tr>
<td>Daily life</td>
<td>drink, stay, hear, lose, family, wait, water, love, leave, pandemic, friend, daily, close, rest, hope, lot, drop, head, play, time</td>
<td>Interesting situation. Got a call from a Mom, Family of 4 lives in a house. Son and her drinks SOULTOX everyday. Daughter and Dad don't. Dad got Covid, then daughter. Mom and Son tested negative 4xs over the 2 wks. Everyone is vaxed. Now she makes everyone drink SOULTOX now.</td>
<td>4.61%</td>
</tr>
<tr>
<td>Political elections</td>
<td>dem, trump, fake, nursinghome, news, pandemic, free, vote, ill, school, election, economy, release, access, truth, guy, forget, sense, deny, hoax</td>
<td>@realDonaldTrump Hey Captain Covid I VOTED FOR JOE BIDEN AND KAMALA HARRIS! The steroids and tranquilizers are making you more batshit crazy than usual. You should go back to the hospital.</td>
<td>3.35%</td>
</tr>
<tr>
<td>Political crisis and discussion</td>
<td>trump, people, person, infect, dangerous, true, covid, condition, shit, happen, stock, send, woman, idea, stupid, destroy, contract, completely, potus, tablet</td>
<td>@realDonaldTrump @USER At least Putin just poisons his political enemies -- Trump wants all Americans to drink poisons such as hydroxychloroquine and oleandrin -- there just aren't enough dead Americans from COVID-19 for Trump -- he wants to kill more with poison -- Trump is a quack ...</td>
<td>3.30%</td>
</tr>
<tr>
<td>Drug scare</td>
<td>country, low, rate, dexamethasone, cost, covid, drug, recovery, black, base, reduce, sell, supply, explain, mortality, expect, increase, panic, produce, improve</td>
<td>Once the pandemic hit, stores really upped the prices of aloe vera leaves like i wouldn’t notice smh</td>
<td>2.87%</td>
</tr>
<tr>
<td>Personal precautions</td>
<td>mask, wear, medicine, social distance, add, avoid, business, plasma, measure, spray, confirm, skin, drive, advise, campaign, practice, oil, wash hand, air, hourly</td>
<td>@USER @USER @USER Try working on your immune system. Covid depletes zinc and vitamin D. Try working from prevention instead of fear.</td>
<td>2.28%</td>
</tr>
<tr>
<td>Public health care</td>
<td>report, support, health, public, supplement, question, kid, child, datum, stage, issue, american, answer, term, prophylactic, mental, phase, safety, concern, inflammation</td>
<td>@USER Then, even hcq, zinc, and zithromax had a estimated 50% success and the CDC and WHO said it didnt, so it shouldnt be used. Imagine half of the death tolls because of at least TRYING something instead of shitcanning it just because you are an &quot;authority&quot; that hates Trump.</td>
<td>2.27%</td>
</tr>
</tbody>
</table>

**Fig. 4** Topic distribution of five top-discussed drugs.
3.4 Co-occurrence networks

We visualized the co-occurrence network for drug-drug and drug-symptom relations in Fig. 5. The nodes represent drugs (Table A.3) or symptoms (Table A.4). Node sizes represent node degrees (i.e., the number of linked entities). Edge weights denote the cosine similarity score of two linked nodes.

3.4.1 Drug-drug network

The origin drug-drug network contained 67 drugs (nodes) with more than 1,000 mentions and 1103 relations (edges) among them. A pre-defined similarity threshold ($\tau$) was established to only visualize relationships with substantial co-occurrence, as measured by cosine similarities exceeding $\tau$. After filtering it with a $\tau$ of 0.005, 62 drugs and 317 relations remained in the network. By using Fast Unfolding (Louvian) algorithm,[35] the drugs were clustered in five categories and were colored in Fig. 5a. The same network with drugs colored by ATC classification (12 types) was shown in Fig. 5b for comparison. Drugs in the same group are denoted with the same color. Both figures share similar clustering characters, especially in psychotropic drugs ATC N (e.g., fentanyl, opium, morphine, etc.) and J (e.g., Lopinavir, Ritonavir, Azithromycin). However, drugs in ATC P group (i.e., ivermectin, hydroxychloroquine, quinine and chloroquine) are clustered with the ATC A group in Fig. 5a. The reason may partially lie in the fact that most parasites are intestinal,[36] so most people who need to take anti-parasitic drugs (i.e., ATC P drugs) often present concomitant digestive manifestations,[37] thus necessitating the use of digestive medications (i.e., ATC A drugs), therefore the two drug groups are closely related. Association between some of the significant drug-drug pairs like two Human Immunodeficiency Virus (HIV) protease inhibitors Ritonavir and Lopinavir has been widely studied.[38] Additionally, through the co-occurrence network, we observed several unusual drug pairings, such as midazolam and morphine, salbutamol and prednisone, and zinc and quinine. These strong co-occurrences suggest potential unexplored
synergistic effects, adverse reactions, or other public health concerns that warrant further investigation. For instance, we noted a distinct correlation between morphine and midazolam, drugs not typically combined in direct COVID-19 treatment. An analysis of all 376 tweets mentioning both drugs revealed that most discussions focused on end-of-life management for COVID-19 patients and on conspiracy theories about the intentional misuse of these drugs, leading to deaths attributed to causes other than COVID-19.

Fig. 5 Visualization of drug-related co-occurrence networks by Gephi, including (a) drug-drug associations based on Gephi clustering (τ=0.005), (b) drug-drug associations based on ATC (τ=0.005) and (c) drug-
symptom associations ($\tau=0.05$). The color dots on the lower right of the figure represent the ATC categories for Fig. 5b.

### 3.4.2 Drug-symptom network

The drug-symptom network had 136 nodes and 3099 edges and was shown in Fig. 5c. After filtering by $\tau$ of 0.05, 50 nodes and 71 edges remained. We observed that the edges often represented symptoms and corresponding treatments, such as Tylenol for fever medication, suggesting the reliability of our association network. We also observed some side effect relations, such as Remdesivir to acute kidney failure[32] and some novel associations receiving no clinical investigation like Molnupiravir to circulatory failure, cocaine to chest cold and vitamin D to malarise. We visualized the top ten closest drugs and symptoms with co-occurrence relationships to the five drugs under investigation (Fig. A.4). These networks revealed the great relevance between hydroxychloroquine, ivermectin and azithromycin from each other. Moreover, Remdesivir was also significantly associated with dexamethasone and Tocilzumab.

### 4. Discussion

This study utilized social media text to develop an NLP-based drug informatic analysis pipeline for assessing public perception of COVID-19-related drugs across time. Leveraging new advances in NLP, we constructed a pretrained language model driven drug entity recognition model and a new targeted sentiment analysis model for polarity prediction of target drugs. This pipeline also includes time trend analysis, topic modeling, and network analysis to explore drug discussions during the pandemic from multiple perspectives. Based on over two years of relevant data, our comprehensive NLP pipeline demonstrates advanced accuracy and completeness in collecting and analyzing data for social media-based drug studies. We will open-source our
pipeline, and it can serve as a comprehensive tool to enhance drug safety control and support public health decision-making after the outbreak of infectious diseases.

Compared to traditional pharmaceutical informatics researches, study of drug-related information on social media exhibits distinctive characteristics and advantages. Social media platforms offer real-time and immediate data, enabling the rapid reflection of drug usage patterns and patient feedback, facilitating the prompt identification of potential risks and benefits.[39-41] Furthermore, social media captures viewpoints and experiences of patients, thus furnishing critical insights for the formulation of patient-centered care.[42, 43] For example, understanding patient’s preference on drugs and disease burden can improve the drug development strategies, enabling pharmaceutical companies to better focus on specific drugs that meet patient needs and preferences.[44] In contrast to previous COVID-19 social media studies,[18, 45-47] this work extracted more rigorous data covering a more extended study period and identified five most discussed drugs to be investigated through a fully data-driven method. The substantial volume of social media data allows for large-scale real-time dynamic analysis, and it also covers a broader population than Electronic Health Records (EHRs), which are confined to hospitalized individuals. Social media datasets could also provide large-scale samples for the detection of rare events and the examination of specific population responses, which are challenge in EHR-based analysis.

Sentiment analysis on drugs can highlight patient misconceptions and disagreements about a specific medication, enabling pharmaceutical companies and public health agencies to address public anxiety and reduce misinformation about drugs. Our results confirmed findings from Hua et al.[18] that the public concern and polarity for ivermectin and hydroxychloroquine, which received most social attention, are highly correlated with emotional and political factors, such as personal political orientation, presidential elections, and conspiracy theories. For instance, there was a surge of approximately 200% in acquisitions of medication
alternatives such as Hydroxychloroquine within two days subsequent to the press briefing conducted by Donald Trump on March 19, 2020.[1] The topic distribution indicated possible effect or side effect of ivermectin on the immune system and the wide in-hospital treatment use of Remdesivir, but the sentiment analysis showed most opposing stances toward Remdesivir which climbed significantly as the crisis unfolded. It was due to shortages, emergency needs, inefficiency, [48] and potential side effects of Remdesivir like bradycardia,[49] increased risk of hepatic, renal and cardiovascular reactions. [50, 51] Some people even hyped up on Twitter that Remdesivir is approved solely for the purposes of reaping big profits for Anthony Fauci and the democidal cabal that he fronts, bilking the tax payer of billions, and all while quietly euthanizing an unwitting public. Moreover, we also found that daily supplements like zinc and vitamin D did not attract much public attention, but their immune-enhancing properties make them significantly more commended by public than the other three drugs, especially Remdesivir.

It holds immense significance for policy makers and public health agencies to timely track and monitor drug-related concerns on social media, especially for new drug candidates during pandemic. Social media swiftly captures drug-related outbreaks and trends, facilitating rapid policy responses for public safety. It offers direct public interaction, enhancing policymaker understanding of public needs and concerns. This aids in shaping policies aligning with public sentiment, boosting policy acceptability and effectiveness.40 Additionally, social media data analysis helps identify drug abuse, adverse reactions, and epidemics, improving health policy planning and resource allocation for evolving health challenges.[52] Our work found that Twitter discussion topics of drugs during COVID-19 were consistent with relevant studies focusing on non-drug COVID-19-related topics.[53-55] Similar to them, this study uncovered public concerns about "public health measures" and "treatment and recovery". In addition, by focusing on drugs, we discovered new drug-specific concerns, such as "drug panic" and "immune response". The focus on "drug panic" may reflect
societal uncertainty and anxiety about the drug use during epidemic. Understanding these anxieties can be instrumental in enabling mental health professionals and policymakers to take measures to support mental health and implementing interventions to alleviate anxiety. Cares about the "immune response" may be indicative of public concerns about the immune system, including vaccines and immunotherapies. This can help health agencies better communicate information about vaccinations and immunization support to increase public awareness of immunization.

Many prior studies aimed to detect potential DDI and ADR from social media[56-58] or online literature[59] but largely depended on external vocabulary for keyword-matching and little visualization was performed. This study utilized advanced pretrained language models to identify drug mentions and classify the corresponding sentiments from social media text, ensures the accuracy of information extraction and sentiment prediction. As the pretrained language model is the main NLP structure in our pipeline, it can be easily extended by integrating better large language models (LLMs) that have a similar deep learning network structure but with larger parameters, given enough computational resources. Visualization module could illustrate associations between drugs, drug-symptoms pairs, and possible clusters or patterns intuitively and clearly, making it’s easier for researchers to understand and interpret the findings for DDI and ADR.[60] In addition, our co-occurrence network analysis found many widely studied drug-drug and drug-symptom pairs which could verify the reliability of network analysis. The clustering results are consistent with the classification of general clinic (i.e., ATC) to a certain extent, suggesting its potential in capturing similarities and associations between drugs. Notably, we also found many drug pairs with not widely examined associations, such as zinc and quercetin. Their complex (Q/Zn) is considered a potential new drug therapy for improving glycemic control and pulmonary dysfunction in diabetes mellitus,[61] which needs to be further investigated. We found new drug-related associations, such as rheumatoid drugs (hydroxychloroquine,
dexamethasone, etc.) may affect COVID-19 treatment due to drug repositioning. Furthermore, networks of top five drugs revealed the significant associations between them such as the co-medication of ivermectin, hydroxychloroquine and azithromycin for COVID-19. Our network analysis also indicated the combination of Remdesivir and Tocilizumab or dexamethasone, and randomized controlled trial showed the efficacy of them for the treatment of severe COVID-19.[62-64]

In essence, the utilization of NLP techniques and network analysis to analyze vast amounts of social media data is an emerging research approach in pharmacovigilance. It holds immense potential in various areas such as the monitoring of ADR, the analysis of drug usage trends, the prediction of epidemics, and the evaluation of drug treatment effects. This novel method has the capacity to serve for pharmaceutical firms, regulatory agencies, and the healthcare fields with more precise and timely information to enhance their efforts in safeguarding public health.

4.1 Limitations

Certain limitations apply to this study. First, social media users can’t represent the general population. For example, Twitter users in the U.S. are younger and more likely to be Democrats, and the most prolific 10% of users create 80% of tweets,[65] which may result in bias of our observations. Secondly, although we tried to automate the information extraction with deep learning, we still relied on an empirical lexicon to cluster different concept representations. This allowed us to effectively reduce false positives but not to avoid false negatives. Thirdly, manual checks for symptom recognition suggested that approximately 2-3% of the tweets may still be false positive (e.g., lexical ambiguity like American fever dream), which would lead fake associations, despite the combination of rigorous rules and advanced NLP models based on deep learning. Data accuracy, as well as the reliability of the analysis, are also limited by the authenticity of social media data and the influence of noisy information.
5. Conclusion

Our study proposed a pipeline of using social media data and NLP techniques to mine potential drug information, timely track drug-related hot events, facilitate public health stakeholders to conduct reasonable policy enactment, monitor drug public opinion and avoid malignant events in the event of a public health emergency. In addition, it can supplement the existing ADR and DDI databases by constructing multiple medical entity co-occurrence networks to provide real-world clues for future research. Our framework applies not only to COVID-19 but also to other periods of epidemics or major social events. It can also target other public health care foci such as vaccination.

6. Summary Table

1. What was already known about the topic?

- The prevalence of the COVID-19 pandemic has induced an urgent demand for efficacious pharmacotherapies.
- Traditional pharmacovigilance mechanisms are plagued by inefficiencies, reporting biases, and a lack of timeliness, thereby lacking comprehensive coverage of the population's sentiments and experiences.

2. What has this study added to our knowledge?

- The utilization of NLP techniques and network analysis to analyze vast amounts of social media data is an emerging research approach in pharmacovigilance.
- Leveraging new advances in NLP, we could address limitations in the former researches, such as high false positives in information retrieval.
- NLP-based pipeline can be a robust tool for large-scale public health monitoring and can offer valuable supplementary data for traditional epidemiological studies concerning DDI and ADR.

Acknowledgements
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Declaration of competing interest**

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.

**Appendix A. Supplementary data**

Supplemental information was submitted in a separate file.

**Author contributions**

J.Y., W.L. and X.X. designed the study. W.L and J.Y. drafted the manuscript. Y.H. collected the data, helped draft and revise the manuscript. W.L. performed data and statistical analysis. P.Z. built the NER and TSA models. L.Z. provided critical reviews. All authors reviewed the manuscript. W.L. takes responsibility for the integrity of the work.

**Data and code availability**

Data, source code, and pipeline tutorial of this paper are available at https://github.com/zju-liwanxin/covid-twitter-drug.

**Reference**


[26] *the Anatomical Therapeutic Chemical Classification System (ATC)* from WHO, <https://www.whocc.no/atc_ddd_index/>. (>)


[58] P.M. Coloma, B. Becker, M.C. Sturkenboom, E.M. van Mulligen, J.A. Kors. Evaluating Social Media Networks in...


Figure Legend

- Fig. 1 Workflow of drug analysis with NLP on Twitter. NLP: natural language process; CT-BERT: COVID-Twitter-BERT; METS-CoV: Medical Entities and Targeted Sentiments on COVID-19-related tweets, NER dataset containing medical entities and targeted sentiments from COVID-19 related tweets.

- Fig. 2 Tweet trends for the top five drugs with the highest occurrences and the case number variation between February 1, 2020 and April 30, 2022.

- Fig. 3 Sentiment analyses of the five top-discussed drug from February 1, 2020 to April 30, 2022, grouped according to their polarity, including (a) sentiment distribution, (b) weekly ratio of positive tweets, and (c) weekly ratio of negative tweets. The denominator of the percentage was the entities with sentiment.

- Fig. 4 Topic distribution of five top-discussed drugs.

- Fig. 5 Visualization of drug-related co-occurrence networks visualized by Gephi, including (a) drug-drug associations ($\tau=0.005$), (b) drug-drug associations based on ATC ($\tau=0.005$) and (c) drug-symptom associations ($\tau=0.05$).

Table Legend

- Table 1 Topic model on drug related Tweets.
Data collection


Remove URLs

169,659,956 COVID-19 related Tweets from 103,682,686 users

Development of an NLP pipeline

Tweets of top 5 drugs: Ivermectin, Hydroxychloroquine, Remdesivir, Zinc, and Vitamin D

2,124,757 Drug-related tweets from 1,800,372 users

CT-BERT finetuned with METS-CoV

Name entity recognition

Target sentiment analysis

Data analysis

1. Time trend analysis
2. Sentiment analysis
3. LDA topic model
4. Co-occurrence network